Faculdade de Engenharia da Universidade do Porto

Mestrado em Engenharia Informática

Anotação Ad-hoc de Conteúdos Audiovisuais

Reutilização de Descritores de Baixo e Alto Nível para Extracção de Conhecimento

Mário Miguel Fernandes Cordeiro
Licenciado em Engenharia Electrotécnica e de Computadores

Dissertação de Mestrado

Porto
Julho de 2008

Anotação Ad-hoc de Conteúdos Audiovisuais
Reutilização de Descritores de Baixo e Alto Nível para Extracção de Conhecimento
Dissertação de Mestrado de
Mário Miguel Fernandes Cordeiro

Realizada sob a supervisão do Professor Doutor
Maria Cristina Ribeiro
Professor Auxiliar
Departamento de Engenharia Informática
Faculdade de Engenharia da Universidade do Porto

Aos meus pais, Mário e Lurdes.

Agradecimentos

À minha orientadora, pela dedicação e paciência demonstrada. Sem a sua ajuda teria sido difícil manter o rumo e chegar a bom porto.

À Arminda e ao Diogo, pelo apoio, pela boa disposição, e por me fazerem sentir dia a dia o quanto é bom termos ao nosso lado as pessoas de quem gostamos.

A Ti por me fazeres acreditar que com fé e dedicação é sempre possível atingirmos os nossos objectivos.

Resumo

Assiste-se actualmente a um investimento significativo em sistemas de recuperação multimédia. A recuperação de imagem baseada em conteúdo (CBIR) é uma área de grande actividade sendo um dos seus objectivos o de diminuir o chamado "fosso semântico" através de sistemas totalmente automatizados para indexação e recuperação. Por outro lado, o acesso generalizado à internet de banda larga está a provocar uma revolução na organização de conteúdos multimédia através de processos de anotação ad-hoc utilizando etiquetas. O método de inserção de etiquetas para descrever conteúdos multimédia embora simples, requer intervenção humana, mas está a transformar-se num processo eficiente e popular para obtenção de metainformação valiosa para a pesquisa. A CBIR está portanto interessada na qualidade semântica dos descritores obtidos através de extracção automática, enquanto a área da anotação está mais preocupada com normas de metainformação que promovam o uso de vocabulários mais uniformes.

O objectivo deste trabalho é mostrar que ambas as abordagens podem ser combinadas num sistema operacional. Propõe-se uma técnica de anotação mista que use quer a similaridade de características de baixo nível quer a recuperação através de anotação textual. Conteúdos multimédia anteriormente anotados podem ser vistos como uma fonte de mapeamento entre características de alto e de baixo nível, e as anotações podem ser propagadas para outros itens multimédia. Com este processo espera-se preservar a qualidade elevada das anotações multimédia manuais, reduzindo o tempo e custo por objecto de vídeo anotado. Os resultados podem ser embebidos em ferramentas que ajudam os utilizadores no processo de anotação, fornecendo sugestões para material multimédia relevante ou relacionado, anteriormente anotado, através de técnicas de recuperação de imagem e vídeo.

A demonstração da abordagem proposta requereu a experimentação sistemática com ferramentas de análise automática de imagem e vídeo, o desenho de uma arquitectura

para o sistema de anotação e uma validação da prova de conceito através de um ambiente experimental. O MPEG-7 Experimentation Model foi adoptado como ferramenta de suporte à extracção de características e à recuperação baseada em similaridade. O conceito de sistema de anotação foi avaliado com 5 colecções distintas de material vídeo variando os descritores e critérios de decisão utilizados.

Abstract

Significant effort is being invested in multimedia retrieval systems. Content-based image retrieval (CBIR) is a particularly active area where one of the intended goals is to bridge the so-called "semantic gap" using fully automated systems for indexing and retrieval. On the other hand, the generalized access to broadband internet is causing a revolution in multimedia content organization with ad-hoc tagging annotation processes. The simple method of inserting tags to describe multimedia content requires human involvement but is becoming an efficient and popular method for obtaining metadata which is invaluable for search. CBIR is therefore interested in the semantic quality of the descriptors which can be obtained by automatic extraction, while the annotation area is concerned with the metadata standards which can favour more uniform vocabularies.

The goal of this work is to show that both approaches can be combined into an operational system. The proposed annotation technique is based on both low-level feature similarity and keyword annotation retrieval. Previously annotated multimedia material is regarded as a source of high-level/low-level feature mappings which can be propagated into other multimedia items. This process is expected to preserve the high quality of manual annotations, reducing the time and cost per annotated video time unit. An annotation tool assists users in the process, giving sugestions to relevant and related previously annotated multimedia material using image and video retrieval techniques.

Demonstrating the proposed approach has required extensive experimentation with automatic image and video analysis tools, the design of an architecture for the annotation system and the validation of the proof of concept using an experimental environment.

The MPEG-7 Experimentation Model has been adopted as a support for feature ex-

traction and similarity retrieval. The annotation concept system was evaluated using 5 distinct video collections using diverse descriptors and decision criteria.

Conteúdo

Lista de Figuras

Lista de Tabelas

Glossário

BiM	*MPEG-7 Binary Format for XML Data*
CBIR	*Content-Based Image Retrieval*
CSAR	*Circular Simultaneous Autoregressive*
CSS	*Curvature Scale Space*
DCT	*Discrete Cosine Transform*
DMS-1	*Descriptive Metadata Scheme - 1*
GIF	*Graphics Interchange Format*
GMRF	*Gaussian-Markov Random Field*
HMHV	*Spectrum Estimation and the Fourier-Mellin transform*
HSV	*Hue, Saturation, Value color model*
HTML	*HyperText Markup Language*
idf	*Inverse Document Frequency*
IPTC	*International Press Telecomunications Council*
JPEG	*Joint Photographic Experts Group*
KLT	*Kanade-Lucas-Tomasi*
LBP	*Local Binary Pattern*
LSI	*Latent Semantic Indexing*
LUV	*Luma Crominance color model*
MIR	*Multimedia Information Retrieval*
MPEG	*Moving Picture Experts Group*
MPEG-1	*MPEG - Audio and Video (AV) Coding and Compression Standard*
MPEG-2	*MPEG - Generic Coding of Moving Pictures and Associated Audio Information Standard*
MPEG-4	*MPEG - Audio and Video Coding Standard and Related Technology*
MPEG-7	*MPEG - Multimedia Content Description Standard*
MPEG-21	*MPEG - Multimedia Framework Standard*
MRF	*Markov Random Field*

MRSAR	*Multiresolution Simultaneous Autorregressive*
MSSG	*MPEG Software Simulation Group*
MXF	*Material eXchange Format*
OWL	*Web Ontology Language*
PNG	*Portable Network Graphics*
PPM	*Portable Pixel Map*
QBIC	*Query By Image Content*
RDF	*Resource Description Framework*
RDFS	*RDF Schema*
RGB	*Red, Green, Blue color model*
SGML	*Standard Generalized Markup Language*
SMAT	*Synchronous Multimedia and Annotation Tool*
SMPTE	*Society of Motion Picture and Television Engineers*
SQL	*Structured Query Language*
tf	*Term Frequency*
tf_idf	*Term Frequency* x *Inverse Document Frequency*
TREC	*Text Retrieval Conference*
TRECVID	*TREC Video Retrieval Evaluation*
VAML	*Video Annotation Markup Language*
W3C	*World Wide Web Consortium*
WAV	*Waveform Audio Format*
MPEG-7 XM	*MPEG-7 eXperimentation Model*
XSD	*XML Schema Definition*
XML	*Extensible Markup Language*

Capítulo 1

Introdução

Com a generalização e massificação das tecnologias digitais a que se assistiu nos últimos anos, tanto as pessoas como as organizações geram, regularmente, grandes volumes de informação multimédia, nas suas diversas formas: imagem, vídeo, som. Encontrar formas de gerir, organizar e recuperar eficientemente documentos na forma digital é importante não só na *web* mas também em ambientes de difusão e distribuição de material multimédia (televisão, rádio, música, fotografia) onde se verifica um grande esforço em sistemas de arquivo, manutenção e reutilização de material multimédia. Ultimamente esta situação estende-se às nossas casas, onde temos também arquivos de material multimédia que cada vez mais queremos catalogar, pesquisar ou até mesmo partilhar. Em suma, diversas áreas beneficiam de resultados em recuperação multimédia.

Um processo de recuperação de conteúdos multimédia tem associadas algumas dificuldades não existentes num processo de recuperação de texto. As técnicas de recuperação de texto são frequentemente baseadas na formulação de interrogações utilizando um determinado conjunto de palavras-chave. Ou seja, as interrogações e os documentos a recuperar partilham a mesma representação e pode optar-se por utilizar directamente os textos dos documentos para os descrever. Em relação a conteúdos multimédia isso já não é aplicável em geral: a formulação de interrogações pode ser feita em diversos níveis de representação que vão desde o totalmente textual até a uma formulação partindo de um outro documento multimédia.

Tomando o exemplo de uma videoteca, a pesquisa de um filme será tanto mais fácil quanto maior for a metainformação adicionada na sua catalogação. A estruturação

aplicando etiquetas como título, género, sinopse, autores, imagem da capa, é fundamental na pesquisa de um filme; a visualização de um excerto, ou uma sequência de imagens estáticas de cenas é uma ajuda preciosa na pesquisa.

Implicitamente neste exemplo de pesquisa falamos de dois paradigmas de pesquisa diferentes. Quando pesquisamos por etiquetas estamos a efectuar uma pesquisa baseada em conceitos de alto nível através de anotações textuais. Quando fazemos uma pesquisa através da visualização de sequências de imagens estáticas ou excertos, recolhemos informação de conteúdo que comparamos mentalmente com o exemplo que procuramos. Um outro aspecto que diferencia estas duas abordagens é o facto de a anotação e catalogação ser só possível com técnicas manuais de inserção de metainformação, enquanto que os sistemas de recuperação baseados em conteúdo tiram grande partido do facto de poderem obter metainformação sobre os conteúdos de forma automática. Este exemplo é elucidativo das duas áreas distintas de investigação aliadas à recuperação multimédia. Na área de anotação e catalogação o objectivo é fornecer bons esquemas de dados e interfaces gráficas que facilitem a recuperação textual. Na área de recuperação de informação baseada em conteúdos o objectivo é a obtenção de bons descritores de conteúdos e bons algoritmos de similaridade para permitir a pesquisa baseada em conteúdos.

1.1 Contexto

Na Enciclopédia Livre (Wikipedia, 2006d) multimédia é definida como "a utilização de diferentes media (ex.: texto, áudio, gráficos, animação, vídeo e interactividade) para expressar informação". A recuperação de informação multimédia (MIR) trata a recuperação de informação em colecções cujos documentos são multimédia. A investigação em recuperação de informação multimédia abrange áreas tais como a pesquisa em texto, a análise de imagem e de vídeo, a indexação e as interfaces com utilizador.

No que respeita à recuperação, embora a definição de multimédia seja lata, existem áreas especializadas no tratamento que cada um dos tipos de media. Enquanto a recuperação de texto está especializada em documentos textuais, a recuperação de imagem utiliza técnicas completamente diferentes da recuperação textual. Neste trabalho o foco será nos problemas colocados pela recuperação em imagem e vídeo.

A recuperação de informação textual assenta em técnicas de análise de texto que, em

conjunto com modelos matemáticos ou estatísticos, conseguem hoje em dia fornecer bons resultados. É um problema que já se encontra bem resolvido, com uma vasta disponibilização de ferramentas de utilização generalizada. Os motores de pesquisa na *web* são os exemplos mais divulgados.

Por outro lado, a recuperação de imagem é ainda uma área em desenvolvimento, sendo a complexidade e custo computacional associados ao processamento de imagem as grandes condicionantes da evolução da recuperação de imagem. A extracção de informação de alto valor semântico é complexa e de difícil automatização, o que faz com que a recuperação de imagem assente na comparação de características de baixo nível de imagens, por exemplo características de cor, forma ou discriminação de objectos. No que respeita às interrogações estas raramente são expressas em características de imagens, usando antes conceitos do domínio dos utilizadores, o que aumenta a complexidade do problema da recuperação. No texto são utilizadas palavras como interrogações, sendo partilhado o nível de representação entre interrogação e documentos. Na imagem a utilização de palavras como interrogação traduz-se numa diferença entre o nível de representação da interrogação e o das imagens que constituem os repositórios (interrogações de texto sobre características das imagens). Há uma clara distinção de nível semântico entre as características das imagens, facilmente extraídas e calculadas, e conceitos de alto nível ou termos intuitivos para os utilizadores. Sebe *et al.* (2003) chama à distância entre esses dois mundos fosso semântico (*semantic gap*).

Em resumo, num sistema de recuperação de imagem a tarefa árdua é a de mapear interrogações de alto nível, especificadas por utilizadores, no domínio dos conceitos de baixo nível onde os resultados podem ser obtidos com maior ou menor número e complexidade de operações de cálculo.

Inicialmente baseada na recuperação de imagem, a recuperação de vídeo é ainda uma área de investigação muito inicial. Do ponto de vista técnico o conteúdo vídeo pode ser considerado como uma sequência de imagens que representam movimento[1]. A recuperação de conteúdos de imagem pode ser feita com grande fidelidade utilizando apenas características de baixo nível, enquanto no vídeo Bertini *et al.* (2002) refere que se torna necessária a disponibilização de informação de mais alto nível para uma boa recuperação. A informação temporal, de que é exemplo o movimento, não pode ser representada apenas numa imagem. Um conhecimento específico do domínio e área de aplicação ajuda à extracção de metainformação de mais alto nível (Colombo *et al.*,

[1] em língua inglesa "moving pictures"

1999), mas por outro lado torna difícil o processo de extracção de conceitos de alto nível por ferramentas de utilização genérica.

No que respeita ao vídeo, o fosso semântico torna-se ainda mais problemático. Passar de imagens para vídeo adiciona ordens de grandeza à complexidade do problema de recuperação devido à necessidade de indexação, análise e pesquisa dos aspectos temporais do vídeo.

Nem as técnicas aplicadas aos tradicionais sistemas de recuperação de informação textual, já de uso generalizado tanto em sistemas privados como na *web* (Vise & Malseed, 2006; Pinkerton, 1994; Yang & Filo, 2006), nem os emergentes sistemas de recuperação de imagem baseados em conteúdo, cujos protótipos disponíveis na *web* fornecem métodos pioneiros de pesquisa em fotografia (flickr, 2006; Cabral, 2006; GIFT, 2006), conseguiram dar uma resposta eficaz aos problemas actuais na área de recuperação de informação vídeo. Actualmente os sistemas de recuperação de vídeo disponibilizados na *web* são totalmente baseados em sistemas de recuperação de texto (Sivic & Zisserman, 2003; Hurley *et al.*, 2006; Herzog *et al.*, 2006) através de descrições introduzidas por utilizadores. Não tiram portanto partido da análise de conteúdos. Por outro lado, e paralelamente às evoluções da *web* no sentido de recuperação de informação, assiste-se à disponibilização de ferramentas de análise e extracção de características de conteúdos que de forma automática fornecem descritores de baixo nível. A correspondência entre conceitos, através de ontologias (Guarino & Welty, 2000), tesauros (Wikipedia, 2007s) e redes de conceitos (Fensel *et al.*, 2001), é também uma área intensa de investigação. A interligação entre conceitos é importante no sentido em que cruza informação obtida através de várias fontes ou técnicas.

Num processo de recuperação de vídeo, como já referido, nem os processos automáticos de extracção de características de conteúdo nem os predominantemente manuais, em que são adicionadas descrições ou anotações aos conteúdos, satisfazem os requisitos de qualidade e escalabilidade. É neste âmbito que será desenvolvido o tema da anotação de vídeo.

1.2 Objectivos

Pretende-se a diminuição ou cruzamento do fosso semântico através de um sistema de anotação baseado em técnicas de recuperação de informação visual que ajude os

utilizadores no processo de adição de metainformação a conteúdos multimédia. A presença de características de baixo nível favorecerá a recuperação baseada em conteúdo, enquanto a metainformação textual associada favorecerá a recuperação através de interrogações textuais. Espera-se também com esta metodologia a diminuição significativa do tempo e esforço empregue na anotação manual de conteúdos vídeo.

Os sistemas de recuperação de informação baseada em conteúdo e os sistemas de anotação seguem actualmente caminhos paralelos. A recuperação baseada em conteúdo está associada a repositórios informais de uso de massas e a formas automáticas de catalogação e recuperação de informação, em que muitas vezes é ignorada por completo a informação de contexto. Baseia-se apenas em descritores extraídos de forma automática e calculados através de características dos conteúdos. Muitas vezes existe informação de contexto importante capaz de diferenciar dois documentos do ponto de vista de análise de características muito semelhantes. Esta informação encontra-se, tipicamente, disponível em descrições ou etiquetas associadas aos conteúdos e quando existente deve ser considerada num processo de recuperação.

Por outro lado, os sistemas de anotação para pesquisa estão mais ligados a repositórios institucionais e utilizam esquemas de descrição normalizados. No caso da anotação em repositórios é dispendido esforço para descrição de conteúdos. Esta metodologia, embora muito utilizada, não tira partido da análise automática de conteúdo, e está por isso muito dependente da metainformação associada aos documentos multimédia, ou seja da metainformação adicionada aquando da anotação.

Há aqui uma dualidade entre as duas abordagens. À recuperação baseada em conteúdo falta tirar partido da possibilidade de adicionara metainformação. A recuperação baseada em normas de descrição fica limitada por ignorar informação de conteúdo. Do ponto de vista de uma melhor eficácia de recuperação surge a necessidade de unificar ambas as abordagens, obtendo uma recuperação baseada em conteúdo e em descrição.

Partindo da importância das descrições e anotações inseridas manualmente pretende-se neste trabalho aumentar a eficácia dos processos de anotação existentes. O objectivo é obter uma a abordagem ao processo de anotação de vídeo através da integração de um sistema de recuperação de informação baseada em conteúdo, facilitando ao utilizador a inserção de metainformação de descrição através da localização de segmentos semelhantes. Trata-se assim de reutilizar a metainformação de alto valor semântico já adicionada, em conteúdos que têm semelhanças do ponto de vista de características de baixo nível.

1.3 Estrutura da Dissertação

No Capítulo 2 são abordados aspectos relacionados com a recuperação de informação. Partindo dos modelos clássicos de recuperação de informação textual e passando pelos sistemas de recuperação de informação visual são abordadas medidas de similaridade entre documentos, tipos de metainformação, uso de ontologias como forma de interligação entre conceitos e processos de interacção com o utilizador. A extracção de características de conteúdo merece uma referência clara neste capítulo uma vez que é nestas características que assentam os sistemas de recuperação baseados em conteúdo. Como complemento são apresentadas referências a sistemas existentes de recuperação de informação baseadas em conteúdo, e metodologias e métricas para avaliação de sistemas.

A anotação de conteúdos multimédia, em particular o vídeo, é o tema base do Capítulo 3, onde são apresentados sistemas de anotação vídeo existentes actualmente, são comparadas as anotações por via manual e automática, e apresentado um conjunto de normas para anotação. É também descrita uma proposta de anotação baseada na reutilização de anotações e cenários de utilização onde essa abordagem é vantajosa.

Como base experimental à proposta de reutilização de anotações, no Capítulo 4 é descrito um sistema de anotação baseado em pesquisa. São abordados os módulos de recuperação e anotação, uma possível arquitectura para um protótipo a ser usado como prova de conceito, e finalmente a descrição detalhada do ambiente experimental montado no âmbito da validação do modelo. Neste capítulo é dedicada uma secção à apresentação de resultados experimentais.

O Capítulo 5 apresenta as conclusões acerca do sistema de anotação baseada em pesquisa, aponta trabalho que pode ser desenvolvido futuramente e dá uma perspectiva acerca de algumas tendências da *web* em relação a sistemas de recuperação multimédia onde as técnicas propostas podem ser vantajosas.

Capítulo 2

Recuperação de Informação Visual

A expressão Recuperação de Informação (*Information Retrieval*) pela primeira vez utilizada por Mooers (1947) capta um conceito multidisciplinar que engloba áreas diversas que vão desde as ciências de computação e tecnologias de informação, até às áreas mais clássicas como as ciências documentais e a biblioteconomia, a psicologia cognitiva, a linguística e a estatística. O termo Informação no sentido mais genérico "(do latim *informatione*) define-se como acto ou efeito de informar ou informar-se; comunicação; indagação, devassa; conjunto de conhecimentos sobre alguém ou alguma coisa; conhecimentos obtidos por alguém; facto ou acontecimento que é levado ao conhecimento de alguém ou de um público através de palavras, sons ou imagens; elemento de conhecimento susceptível de ser transmitido e conservado graças a um suporte e um código" (Wikipedia, 2006c). Por outro lado, a informação hoje em dia remete-nos para os meios divulgação de informação, isto é jornais, notícias, televisão. Fazendo a transposição para as novas tecnologias de comunicação a informação é feita utilizando meios mais modernos como é o caso de computadores pessoais e internet, áudio e vídeo, telefones. Sendo assim, informação é "termo que designa o conteúdo de tudo aquilo que trocamos com o mundo exterior e que faz com que nos ajustemos a ele de forma perceptível" (Wikipedia, 2006c). Conclui-se que a palavra Informação é um termo de significado lato cuja fronteira é de difícil delimitação.

Num contexto de Recuperação de Informação Weaver & Shannon (1949), na sua abordagem à teoria da comunicação, descreve tecnicamente o conceito de informação como algo que não é directamente calculado. Ou seja segundo o conceito mais generalizado de Recuperação de Informação seria mais correcta a substituição do termo Informação

por Documento. Mesmo assim o termo Recuperação de Informação é comummente utilizado com o significado de recuperação de documentos por autores como Cleverdon (1970), Salton (1970), Jones (1968) e Lancaster (1968). Uma bom exemplo dessa definição é a de Lancaster (1968): "Um sistema de recuperação de informação não informa, isto é, modifica o conhecimento do utilizador acerca do assunto objecto de pesquisa. Apenas o informa acerca da existência (ou não existência) e localização de documentos relacionados com o seu pedido."

Essa definição de Recuperação de Documentos será também a adoptada no âmbito desta dissertação. Deve ainda fazer-se menção ao tipo de estruturação da informação a recuperar. O termo Recuperação de Informação refere-se também à recuperação de documentos não estruturados, isto é, documentos cujo conteúdo se encontra sob a forma de texto em linguagem natural. Outro tipo de documentos não estruturados também são considerados em várias áreas da Recuperação de Informação, nomeadamente a Recuperação de Informação Visual: imagens fotográficas, áudio, vídeo. O foco do trabalho é a recuperação de documentos multimédia, em particular, documentos vídeo.

Quando se fala de recuperação de documentos multimédia, fala-se na realidade de um conjunto de documentos muito distintos na forma, estruturação e próprio conteúdo. Nas Secções 2.1 e 2.2 são abordados dois modelos distintos de recuperação multimédia: a recuperação de informação textual e a recuperação de informação visual. A abordagem a estas duas vertentes distintas da recuperação coloca questões relacionadas com a metainformação (*metadata*[1]) retirada da análise dos conteúdos. A Secção 2.3 expõe os vários tipos de metainformação que podem retirados de conteúdos multimédia. No caso dos sistemas de recuperação de imagem a extracção de informação dos conteúdos requer técnicas mais elaboradas e a Secção 2.4 aborda alguns dos tipos de características e técnicas de extracção que podem ser aí aplicadas. As Secções 2.5, 2.6 e 2.7 abordam respectivamente a utilização de ontologias, técnicas de interacção com utilizador utilizadas para aumentar as taxas de recuperação, e apresentação de sistemas existentes na actualidade de recuperação de informação visual. Por último, a Secção 2.8 dedica-se exclusivamente à apresentação de métricas de avaliação de desempenho de sistemas de recuperação.

[1] *the bits about the bits* (Nack & Lindsay, 1999a,b)

2.1 Recuperação de Informação Textual

O objectivo de um Sistema de Recuperação de Informação é, através de pesquisas de termos ou conjuntos de termos inseridos por um utilizador, recuperar documentos relevantes para essa interrogação. Esses termos são comparados, utilizando estratégias próprias, com os documentos previamente armazenados.

No projecto ou desenvolvimento de um sistema deste tipo, há várias técnicas que podem ser utilizadas considerando os esquemas de representação dos documentos, a formulação da interrogação (*query*) e a elaboração de uma estratégia de ordenação (*ranking*), isto é, a métrica de relevância dos documentos em relação à consulta. De entre os vários modelos propostos (Korfhage, 1997; Wong & Yao, 1995) salientam-se o modelo booleano, o modelo de espaço vectorial, o modelo de vector de contexto, o modelo de indexação por semântica latente (LSI), o modelo difuso e o modelo probabilístico. As secções seguintes introduzem cada um destes modelos.

2.1.1 Modelo Booleano

O modelo booleano, um dos mais simples dos modelos clássicos, é ainda utilizado na actualidade. Este modelo baseia-se na álgebra booleana para a realização de pesquisas sobre sistemas de informação fazendo uso dos operadores lógicos AND, OR e NOT para especificar as interrogações (Witten *et al.*, 1999). Partindo de um conjunto de termos ligados por operadores lógicos, são recuperados os documentos que satisfazem a expressão booleana correspondente (Korfhage, 1997).

Este modelo pode ser utilizado em pesquisas sobre arquivos de texto, embora seja mais adequado à pesquisa em dados estruturados como é o caso das bases de dados relacionais. Neste caso as interrogações são feitas com recurso a uma linguagem própria: SQL (*Structured Query Language*) (Wikipedia, 2007r).

A utilização do modelo booleano não permite estabelecer métricas de ordenação por relevância dos documentos recuperados, uma vez que todos os documentos que satisfaçam a condição booleana serão considerados igualmente relevantes para a interrogação (Wong & Yao, 1995).

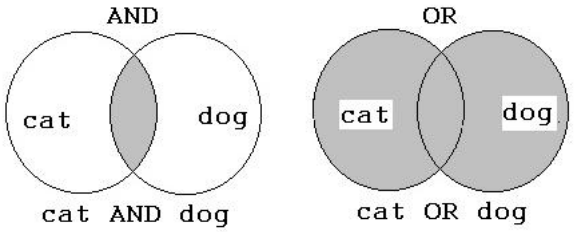


Figura 2.1: Exemplo de operadores AND e OR no Modelo Booleano. Fonte: Wikipedia (2007c)

2.1.2 Modelo de Espaço Vectorial

Os componentes de um sistema de recuperação (documentos, interrogações) são modelados como elementos de um espaço vectorial (Wong & Raghavan, 1984; Dominich *et al.*, 2000; Carleton *et al.*, 1995).

Neste modelo, a resposta a uma interrogação é obtida através da determinação da similaridade entre os vectores dos documentos e o vector da interrogação. Na construção desses vectores há que ter em consideração dois factores: a forma como, partindo dos documentos, são extraídas as características para construção dos vectores; e a redução de dimensionalidade e normalização dos termos.

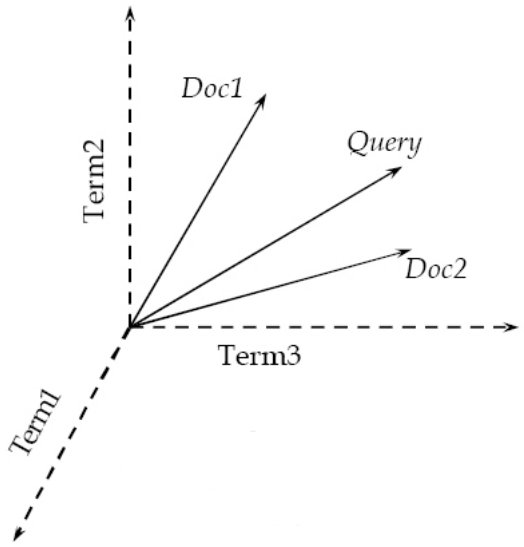


Figura 2.2: Exemplo de modelo de espaço vectorial (documentos, termos e interrogação). Fonte: Wikipedia (2007t)

No modelo de Espaço Vectorial é comum a utilização de pesos, tipicamente calcula-

dos com base em frequências normalizadas para os termos presentes no documento. Os termos com número de ocorrências maior podem assim ter pesos maiores. Numa interrogação normal, todos os termos tomam pesos iguais, podendo ser alterados seguindo preferências dos utilizadores.

A comparação entre dois vectores, num espaço vectorial, pode ser feita através da medida do co-seno do ângulo que formam (Korfhage, 1997; Arfken, 1985). Vectores com um ângulo próximo de 0 graus correspondem a documentos com uma elevada similaridade. O modelo vectorial permite o cálculo de uma métrica de ordenação de similaridades baseada nesta medida. Outras formas para cálculo de medidas de similaridade são a medida do pseudo co-seno (Pascasio & Terwilliger, 2003), a medida de Dice (Salton, 1989), as medidas de correlação (Wikipedia, 2007e) e de co-variância (Wikipedia, 2007f).

Considerando o modelo do espaço vectorial, os documentos relevantes para uma determinada consulta são aqueles representados por vectores próximos do vector que representa a interrogação.

A medida de similaridade através do cálculo do co-seno é muito utilizada no modelo de recuperação de informação vectorial devido à sua estabilidade. Segundo Chiao & Zweigenbaum (2002) esta medida é também a que apresenta o melhores resultados comparada com outros métodos de cálculo de distâncias entre vectores.

A fórmula utilizada para a medida do co-seno é:

$$\cos(\vec{q}, \vec{d}) = \frac{\sum_{i=1}^{n} q_i d_i}{\sqrt{\sum_{j=1}^{n} q_i^2} \sqrt{\sum_{j=1}^{n} d_i^2}} \tag{2.1}$$

onde $\vec{q}$ e $\vec{d}$ são os vectores de interrogação e documento respectivamente. Esta medida valoriza a ocorrência simultânea de termos (representados pelas componentes homologas q_i e d_i) na interrogação e nos documentos e é normalizada pela medida euclidiana dos vectores. Quando dois vectores que são comparados possuem os mesmos termos e pesos, ou seja, são idênticos, o ângulo entre eles é zero e segundo a equação do co-seno tomará o seu valor máximo, ou seja 1. Quanto mais próximo de 1, maior similaridade há entre os vectores.

A representação de documentos por vectores requer uma medida de peso dos termos, sendo comummente utilizada a medida denominada *tf_idf* (*Term Frequency* x *Inverse*

Document Frequency), que considera a frequência intra-documento (frequência do termo no documento) e a frequência inter-documentos, ou seja, a percentagem de documentos da colecção em que o termo ocorre.

Para cálculo da frequência intra-documento (*tf*), pode realizar-se uma normalização das frequências de cada termo no documento com relação à frequência do termo com maior número de ocorrências. A frequência intra-documento é assim definido por

$$tf_i = \frac{f_i}{max(f_d)} \tag{2.2}$$

onde f_i é a frequência para cada termo i e $max(f_d)$ é a maior frequência no documento d.

A frequência inversa do documento (*idf*) é definida como

$$IDF_i = \log\left(\frac{N}{n_i}\right) \tag{2.3}$$

onde N é o número de documentos na colecção e n_i é o número de documentos na colecção em que ocorre o termo i.

O cálculo de *tf_idf* usa o produto destas duas grandezas e alguns factores determinados empiricamente.

$$tf\_idf_i = tf_i \times \log\left(\frac{N}{n_i}\right) \tag{2.4}$$

2.1.3 Modelo Probabilístico

O modelo probabilístico utiliza pesos binários que representam a presença ou não de termos nos documentos. O vector de resultados obtido pelo modelo baseia-se no cálculo da probabilidade de um documento ser relevante para a interrogação. Matematicamente este modelo usa o teorema de Bayes (van Rijsbergen, 1979) para relacionar as probabilidades de relação de termos e documentos. O Princípio de Ordenação por Probabilidade (*Probability Ranking Principle*) descrito por Robertson (1997) serve de base a este modelo e indica que a relevância de documento para uma determinada interrogação é independente de outros documentos.

Para o modelo probabilístico, os pesos dos termos são todos binários, $w_{i,j} \in \{0,1\}$, $w_{i,q} \in \{0,1\}$. A interrogação q é um subconjunto dos termos dos índices. Considerando R o conjunto dos documentos relevantes conhecidos e $\overline{R}$ o complemento de R, ou seja o conjunto dos documentos não relevantes, $P(R|\overrightarrow{d_j})$ representa a probabilidade do documento d_j ser relevante à interrogação q e $P(\overline{R}|\overrightarrow{d_j})$ a probabilidade de ser não relevante. A similaridade do documento d_j em relação à interrogação q é definida por

$$sim(d_j, q) = \frac{P(R|\overrightarrow{d_j})}{P(\overline{R}|\overrightarrow{d_j})} \tag{2.5}$$

que através da regra de Bayes e se for suposta a independência de termos, pode ser aproximada por

$$sim(d_j, q) \sim \sum_{i=1}^{t} w_{i,q} \times w_{i,q} \times \left(\log \frac{P(K_i|R)}{1 - P(K_i|R)} + \log \frac{1 - P(K_i|\overline{R})}{P(K_i|\overline{R})} \right) \tag{2.6}$$

$P(K_i|R)$ indica a probabilidade de o termo representado pelo índice K_i estar presente num documento obtido aleatoriamente do universo de documentos R. Do mesmo modo que $P(K_i|\overline{R})$ indica a probabilidade desse mesmo termo não estar presente no documento. Essas probabilidades só são possíveis de calcular conhecendo à priori o conjunto dos documentos relevantes e não relevantes para os termos indicados. Para contornar essa situação o método é utilizado recursivamente considerando valores iniciais aproximados para $P(K_i|R)$ e $P(K_i|\overline{R})$.

O modelo probabilístico tem como vantagem principal a ordenação por ordem decrescente de probabilidade de relevância dos documentos. Algumas desvantagens são a necessidade de estimar inicialmente a separação dos documentos em conjuntos de documentos não relevantes e relevantes, o facto do método não tomar em conta a frequência com que os termos ocorrem nos documentos (os pesos utilizados são binários) e a suposição de que existe independência de termos.

2.1.4 Modelo de Vector de Contexto

O modelo de vector de contexto é uma extensão ao modelo de espaço vectorial. No modelo de espaço vectorial, os vectores são compostos utilizando apenas a listagem dos

termos e suas frequências de ocorrência. Assim o espaço vectorial representa termos e frequências para cálculo de similaridades com outros vectores. A insensibilidade aos relacionamentos semânticos existente entre os termos incluídos na interrogação e os termos existentes nos documentos é a grande limitação do modelo de espaço vectorial. Ou seja, documentos contendo termos semanticamente semelhantes aos da interrogação podem ficar excluídos do resultado (Billhardt *et al.*, 2002). O modelo de espaço vectorial foi expandido para agregar aos termos do vector o contexto em que ele está inserido, ou seja, outros termos que podem indicar alguma relação semântica.

Schutze (1992) considera o significado semântico dos termos e seus contextos como vectores de um espaço vectorial em que as dimensões correspondem aos termos. Esses vectores são denominados vectores de contexto por possuírem o contexto onde cada termo está inserido no documento. Este contexto é obtido através dos termos próximos, dentro do texto, considerando uma janela de termos que indica quantos deles antes e depois serão considerados na definição do contexto. Assim, para cada termo pode ser gerado um vector contendo os termos próximos do contexto. O vector de contexto de um documento é formado por esses vectores de contexto relacionados com cada termo (Caid & Carleton, 1994).

2.1.5 Modelo de Indexação por Semântica Latente

A maioria dos métodos considera a ocorrência dos termos (termos que constituem a interrogação) nos documentos para realização de cálculos de similaridade que indicarão o grau de relevância de um documento relativamente a uma interrogação. A desvantagem desta abordagem reside no facto de que documentos, cujo grau de relevância é elevado mas que não contêm nenhum dos termos especificados na interrogação, não são apresentados no resultado da pesquisa. Num processo de recuperação alguns dos termos chave dos documentos podem ser esquecidos ou desconhecidos do utilizador que constrói a interrogação. Deste modo resulta a não recuperação de alguns documentos relevantes.

O vocabulário utilizado pelos humanos é extenso e caracteriza-se por uma utilização frequente de sinónimos. A comparação textual entre termos pode ser limitativa para a recuperação de informação baseada no seu significado. O modelo de Indexação por Semântica Latente (LSI - *Latent Semantic Indexing*) (Furnas *et al.*, 1988) utiliza uma abordagem que toma em consideração co-ocorrências de termos, sendo uma alternativa

eficaz na resolução deste problema. A co-ocorrência de termos verifica-se para conjuntos de termos que são encontrados com frequência nos mesmos documentos. Se um determinado conjunto de termos surge com elevada frequência em diversos documentos relativos a uma determinada área de conhecimento ou tema, significa que pode haver uma relação de semântica latente entre esses termos, ou seja uma relação não explícita. Utilizando técnicas estatísticas, o modelo de indexação por semântica latente pode encontrar (ou evidenciar) as possíveis correlações existentes entre documentos (Manning & Schtze, 1999). A Figura 2.3 ilustra esse processo.

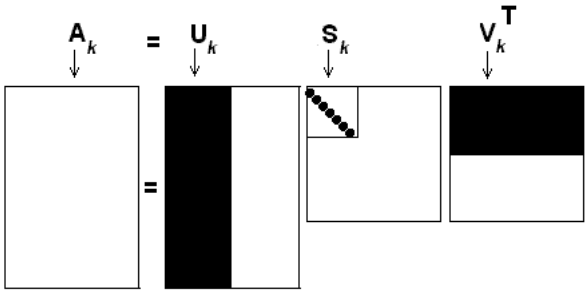


Figura 2.3: Indexação por semântica latente aplicada a uma matriz de termos e documentos. Fonte: Wikipedia (2007m)

2.1.6 Modelo de Lógica Difusa

O modelo de Lógica Difusa é baseado na Teoria dos Conjuntos Difusos (Zadeh, 1965) e define a pertença de um elemento a um conjunto generalizando a função de pertença do intervalo fechado e discreto $[0, 1] \in \mathcal{I}$ para o intervalo contínuo $[0, 1] \in \mathcal{R}$. Sendo assim, define-se um conjunto difuso por $x \in \overline{A} \mid \mu_x(A)$, em que $\overline{A}$ é o universo dos elementos, μ_x a função de pertença e $[0, 1]$ é intervalo em $\mathcal{R}$.

Os modelos baseados na Teoria dos Conjuntos puderam ser generalizados para a Teoria dos Conjuntos Difusos, sendo a pesquisa difusa uma generalização da pesquisa usando o modelo booleano. A lógica difusa fundamenta sistemas de raciocínio aproximado e utiliza os graus de pertença nos conjuntos.

Ao contrário do modelo booleano em que a função de pertença toma apenas os dois valores verdadeiro ou falso (um termo está incluído ou não num documento), no modelo difuso a função de pertença é dada por uma atribuição de pesos (um termo está relacionado com um determinado documento com um peso X enquanto que outro está

relacionado com um peso Y) (Wikipedia, 2007h,g). Esta atribuição de pesos faz com que o conjunto de termos para documentos seja um conjunto difuso. Para passar de um conjunto difuso para um conjunto preciso basta fixar um limiar para a função de pertença. A Figura 2.4 ilustra esse cenário.

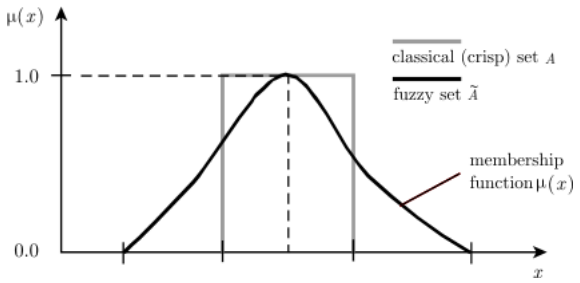


Figura 2.4: Conjunto difuso e limiar para modelo booleano. Fonte: Wikipedia (2007i)

O Modelo de Lógica Difusa é utilizado para captar a relação entre termos e documentos de forma mais fina que o Modelo Booleano. No processo de recuperação faz uso de todo o cálculo associado à Lógica Difusa.

2.2 Recuperação de Informação Visual

Uma das técnicas mais utilizada na área de recuperação de informação visual é a recuperação de informação baseada em conteúdo: QBIC[2] (Niblack *et al.*, 1993; Smith & Chang, 1999).

Marsico *et al.* (1997) referem diversas metodologias a ser aplicadas num processo de recuperação de informação baseada em conteúdo: pre-processamento, segmentação, extracção de características, indexação e recuperação das imagens propriamente ditas. Alternativamente Niblack *et al.* (1993), Smith & Chang (1999) e Ogle & Stonebraker (1995) abordam o problema como um todo. Em ambas as situações o processo de QBIC é realizado em duas partes, uma *off-line*, onde são gerados os índices de cada uma das imagens de entrada, e outra *on-line*, onde são efectuadas as pesquisas propriamente ditas.

A indexação é dependente do tipo de característica que se pretende extrair e é um processo executado normalmente *off-line*. A eficácia do processo de recuperação de

[2]do Inglês *Query By Image Content*

informação está intrinsecamente relacionado com o tipo de características a extrair dependendo do tipo de imagens que se pretende indexar. A extracção de características é normalmente pesada do ponto de vista computacional, exige um elevado poder de processamento, sendo também um processo moroso.

A pesquisa de imagens é realizada normalmente através de uma imagem exemplo, de um esboço ou através de comparação de características predeterminadas da imagem. Para cálculo das similaridades entre as imagens são usadas métricas de distâncias: entre os índices das duas imagens, entre os índices do esboço e da imagem, ou através de características das duas imagens que fornecem a medida de similaridade. A ordenação dessas medidas por similaridade fornece uma listagem das imagens relevantes para a pesquisa. Este processamento obviamente deverá ser efectuado no modo *on-line*.

Nos sistemas de recuperação de informação visual, o processo de extracção de características é sem dúvida a área que tem tido uma investigação mais diversificada, tanto na escolha adequada das características ideais para representação visual, como na forma como podem ser extraídas, e mesmo pela melhor metodologia de armazenamento. No âmbito do trabalho desenvolvido até hoje essas características conseguem agrupar-se em quatro tipos: cor, forma, textura e relacionamento espacial, entre objectos ou entre regiões da imagem. No caso de recuperação de informação sobre repositórios vídeo, a informação temporal é importante para a recuperação.

2.3 Tipos de Metainformação

Rowe *et al.* (1994) caracterizaram e identificaram não só os tipos interrogações possíveis sobre documentos multimédia, como também, os tipos de índices necessários para a concretização dessas interrogações. Esses índices recaem sobre três categorias de dados:

- **Metainformação Bibliográfica** - São dados não directamente relacionados com o conteúdo multimédia em si, mas que fornecem informação adicional e complementar acerca deste. Esta categoria inclui informação acerca do documento (ex.: título, resumo, assunto, género) e indivíduos envolvidos na sua criação (ex.: produtor, realizador, elenco);
- **Metainformação Estrutural** - Esta categoria pode ser descrita, para o caso de

um documento multimédia vídeo, com recurso à hierarquia existente de filme, segmento, cena e toma. Cada uma destas entradas pode ser decomposto em uma ou mais entradas do nível inferior (ex.: um segmento é composto por uma sequência de cenas e uma cena por uma sequência de tomas) (Davenport *et al.*, 1991);

- **Metainformação de Conteúdo** - Representa a informação acerca do conteúdo em si, sendo particularmente importante nos conteúdos audiovisuais. Devido à natureza do vídeo o conteúdo visual é uma combinação de conteúdo estático (*frames*) e conteúdo dinâmico (informação temporal). Os índices podem ser conjuntos de imagens que representam amostras do vídeo, índices de palavras-chave retiradas de diálogos ou som e índices de objectos que representam os momentos de entrada e saída de objectos ou indivíduos na duração da sequência.

2.4 Extracção de Características

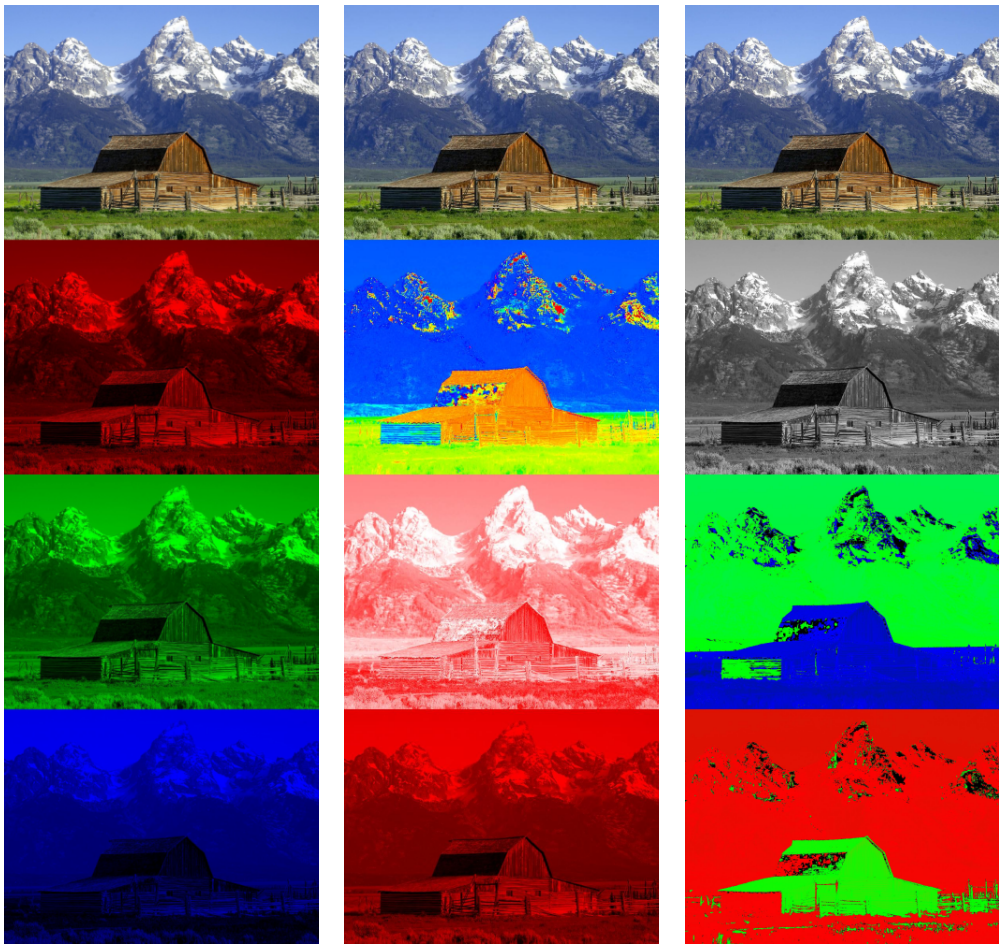
Para que seja possível a recuperação de informação baseada em conteúdo é necessária a utilização de representações dos conteúdos que possam servir de base à comparação entre interrogações e documentos. Nas secções seguintes serão abordadas tipos de características de baixo nível que podem ser extraídas dos conteúdos de imagem e vídeo. Também são mencionados os respectivos descritores MPEG-7 (Martínez, 2002) normalizados para características de Cor, Forma, Textura e Movimento.

2.4.1 Características de Cor

No âmbito da recuperação de informação visual é corrente a utilização de características de cor. Colombo & Bimbo (1999) refere a cor como sendo a característica mais utilizada pelos seres humanos para reconhecimento e discriminação visual. A informação de contexto como auxílio à descrição de cor é muitas vezes utilizada para descrever certas cores base; exemplos comuns são o vermelho Ferrari e o azul bebé. Num processo de extracção automática de características de cor, realizado por computadores, essa referência de contexto é inexistente o que dificulta, por vezes, a distinção entre informação de cor e distorção de cor, isto é cor efectiva e cor observada. A aparência das cores no mundo real é geralmente alterada pela superfície, iluminação/sombra de

outros objectos, e condições de observação e captura o que dificulta ainda mais a sua discriminação sem essa informação de contexto.

Na recuperação de informação visual a indexação de características de cor é feita recorrendo a histogramas de cor. O cálculo desses histogramas é feito através de operações simples que indicam o número de ocorrências de cada cor. Para cada canal do espaço de cor (RGB, HSV, LUV) é feita uma discretização das ocorrências em intervalos. A escolha de um número reduzido de intervalos aumenta a eficácia da computação, reduzindo a exactidão da representação da imagem.



Componentes RGB Componentes HSV Componentes LUV

Figura 2.5: Sistemas de cor RGB, HSV e LUV. Fonte: Wikipedia (2007q,k,x)

Além de não necessitarem de algoritmos computacionalmente pesados, os histogramas de cor têm propriedades de invariância à escala, à translação e à rotação, simplificando a identificação de imagens com propriedades de cor semelhante. A maior desvantagem prende-se com o fraco relacionamento espacial entre objectos ou regiões das

imagens: cenas semanticamente diferentes podem possuir histogramas de cor iguais; cenas semanticamente iguais podem apresentar histogramas de cor diferentes. Este facto torna difícil a distinção de objectos em imagens sem recurso a outras técnicas (Zhang & Lu, 2003; Ojala *et al.*, 2002a,b). A diminuição do número de níveis de quantificação de cor utilizados tende a agravar ainda mais este problema.

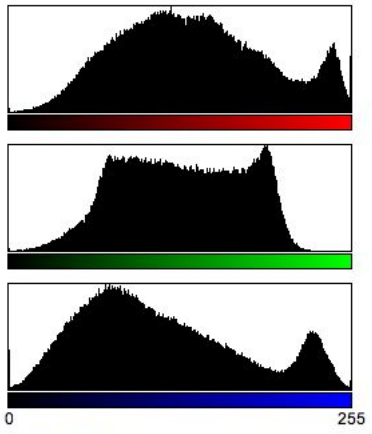


Figura 2.6: Histograma de cor RGB. Fonte: Wikipedia (2007d)

Huang *et al.* (1997) propuseram a utilização de correlogramas relacionando as cores do pixels com as distâncias entre eles de forma a colmatar este problema. Ojala *et al.* (2001) propõe a utilização do espaço de cor HSV, em alternativa ao RGB. Os resultados obtidos demonstraram algumas melhorias na distinção da categoria semântica em que a imagem havia sido enquadrada manualmente. Supostamente, o espaço de cor HSV, fornece uma melhor percepção humana da diferença entre cores do que outros espaços de cor, nomeadamente o RGB. Tao & Grosky (2001) propuseram a utilização de anglogramas de cor, que são histogramas dos ângulos formados pelas posições de regiões semelhantes de cor. Os anglogramas de cor conservam as três propriedades dos histogramas de cor (invariância à escala, translação e rotação).

Todas as características de cor apresentadas são discretas e de dimensão constante, ou seja o número de atributos numéricos da característica é constante e independentemente do tamanho da imagem a que se refere. São portanto utilizáveis na comparação vectorial.

Descritores MPEG-7 de Cor

- *Color Layout* - Especifica a distribuição espacial de cor posteriormente utilizada

para a recuperação ou pesquisa rápida de imagens semelhanças na distribuição de cor. O descritor é extraído de uma matriz 8x8 de cores dominantes previamente calculadas dividindo a imagem em 64 blocos e calculando a cor dominante para cada um deles. Os descritores são comparados através das matrizes de cor dominante.

- *Color Structure* - Captura a informação de cor através de um histograma de cor e informação espacial da imagem. A organização espacial das cores na sua vizinhança é determinada através de um elemento quadrado. O seu tamanho é relativo à dimensão da imagem. O descritor produz um histograma que pode ser comparado com outros seguindo a normalização L1 (Horn & Johnson, 1990).

- *Dominant Color* - Especifica um conjunto de cores dominantes numa região arbitrária. A quantização de cor usando o *Generalized Lloyd Algoritm* (Lloyd, 1982) no espaço de cor LUV é utilizada para extrair um pequeno número de cores representativas em cada região da imagem. Os descritores são comparados com uma medida de coerência espacial. Este descritor é bom para representar características locais (objectos ou regiões de imagens), onde um pequeno número de cores é suficiente para caracterizar o conteúdo de cor.

- *Scalable Color* - É um histograma de cor no espaço de cor HSV, que é quantificado uniformemente em 256 intervalos de acordo com as tabelas fornecidas na norma MPEG-7. Os valores do histograma são quantificados não linearmente através da transformada de Haar (Haar, 1910) até ao nível pretendido. Os descritores podem ser comparados através da reconstrução do histograma, no domínio dos coeficientes de Haar (Haar, 1910), ou através da distância de Hamming (Exoo, 2003) dos sinais dos coeficientes.

2.4.2 Características de Forma

A forma é um critério importante na identificação de objectos tendo em consideração o seu perfil e estrutura física. A sua utilização está bastante vulgarizada em sistemas de análise de imagem onde os objectos têm muitas semelhanças de cor e textura, como por exemplo na imagem médica (Gonzalez & Woods, 2001). A utilização de características de forma é portando a alternativa mais eficaz para fazer a sua diferenciação.

A literatura refere duas metodologias de representação objectos através da sua forma: técnicas baseadas em regiões e técnicas baseadas em contornos.

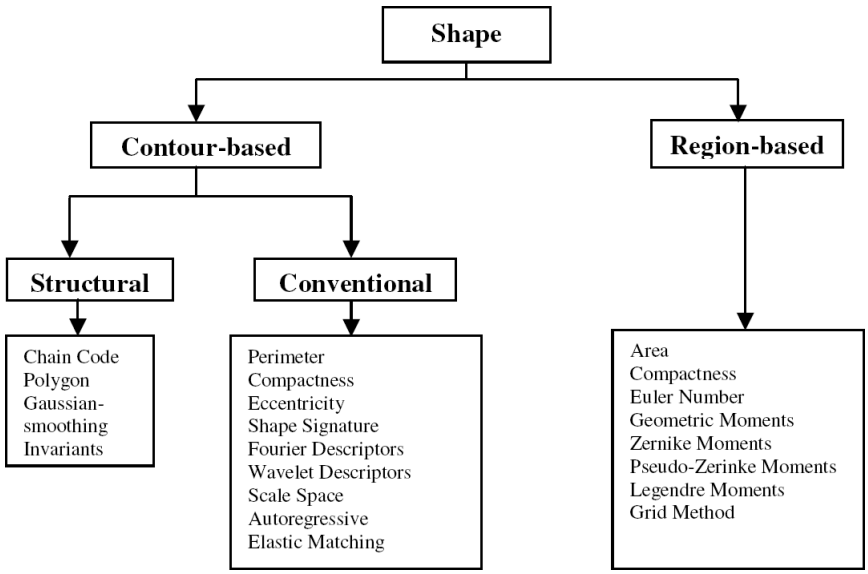


Figura 2.7: Taxonomia dos descritores de forma. Fonte: Zhang & Lu (2003)

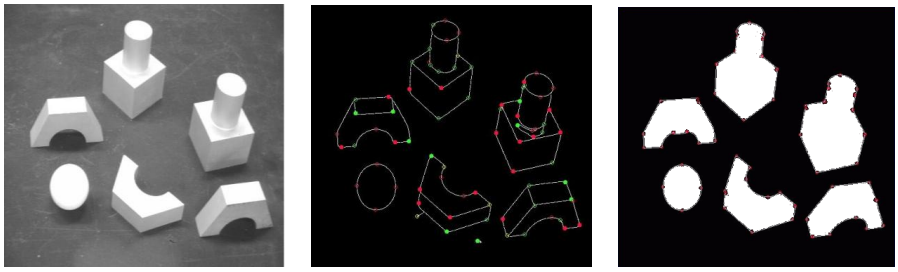
Nas técnicas baseadas em regiões, todos os pixels (*pixels*) são tomados em conta na representação da forma. Geralmente são utilizados momentos para descrever as formas (Niblack *et al.*, 1993; Hu, 1962; Liao & Pawlak, 1996; Teh & Chin, 1988; Taubin, 1992). Dentro deste tipo encontram-se os momentos centrais (Kenney & Keeping, 1951), momentos de Legendre (Teh & Chin, 1988), momentos de Zernike (Teh & Chin, 1988) e momentos de pseudo-Zernike (Teh & Chin, 1988). A utilização de grelhas de representação é também feito em algumas aplicações (Chakrabarti *et al.*, 2000; Lu & Sajjanhar, 1999; Safar *et al.*, 2000).

As características de contornos subdividem-se ainda em Estruturais e Convencionais. Nas abordagens convencionais o contorno é tratado como um todo, ou seja, é mapeado num vector que o descreve. A forma de cálculo de similaridade é normalmente a distância Euclidiana entre os dois vectores. As características Estruturais subdividem o contorno em segmentos (também apelidados de primitivas) utilizando determinados critérios. A representação final é normalmente uma *string* ou uma árvore. A medida de similaridade é calculada através de comparação directa entre as *strings* ou ramos da árvore.

Representações convencionais de forma incluem os descritores globais de forma (*global shape descriptors*) (Niblack *et al.*, 1993), assinaturas de forma (*shape signatures*) (Davies,

2004), descritores espectrais (*spectral descriptors*) (Persoon & Fu, 1986; Kauppinen *et al.*, 1995), espaço de curvatura de escala (*curvature scale space*) (Mokhtarian *et al.*, 1996), comparação elástica (*elastic matching*) (Bimbo & Pala, 1997) e método autoregressivo (*autorregressive method*) (Kauppinen *et al.*, 1995).

As representações Estruturais da forma incluem métodos de aproximação poligonais (Grosky *et al.*, 1992; Mehrotra & Gary, 1995) e cálculo de invariantes de forma (Huang & Huang, 1998; Li, 1999).



Original Contorno Regiões

Figura 2.8: Extracção de formas baseadas em contorno e regiões. Fonte: Zheng & Gao (2004)

Descritores MPEG-7 de Forma

- *Fourier Descriptor* - Este descritor é obtido através da aplicação de uma Transformada de *Fourier* num tipo de assinatura de forma (*shape signature*), ou seja é um descritor baseado em contornos no domínio das frequências. O conjunto de coeficientes da Transformada de *Fourier* é apelidado de descritor de *Fourier* da forma. A assinatura de forma é uma função unidimensional derivada a partir das coordenadas da fronteira da forma. Diversos tipos de assinaturas de forma podem ser utilizados no cálculo do Descritor de *Fourier*. As coordenadas complexas, função de curvatura, função cumulativa de ângulos e função de distância centróide são alguns dos mais comuns. Zhang & Lu (2002) refere que o descritor derivado da função de distância centróide é mais efectivo que um descritor derivado de outra qualquer assinatura de forma.

- *Curvature Scale Space Descriptor* - Este descritor é baseado na análise de contorno considerando a fronteira da forma como um sinal unidimensional. A este sinal unidimensional dá-se o nome de função de curvatura no espaço de escala (*Space Scale*). As passagens por zero são examinadas às diferentes escalas com o intuito

de encontrar as concavidades e convexidades que são úteis no âmbito da descrição de uma forma uma vez que capturam as características perceptivas de um contorno de uma forma.

- *Geometric Moment Descriptor* - É um descritor baseado em regiões e faz uso de momentos invariantes como método de representação de uma forma. Os momentos invariantes são derivados dos momentos de forma e são invariantes a transformações geométricas em duas dimensões. Este descritor é muito compacto e não é exigente do ponto de vista computacional.

- *Zernike Moment Descriptor* - Sendo também um descritor baseado em regiões faz uso de momentos ortogonais para recuperar a imagem de momentos baseados na teoria de polígono ortogonais. Os momentos de Zernike (Teh & Chin, 1988) permitem a construção de momentos invariantes para uma ordem arbitrária.

- *Grid Descriptor* - Este descritor pode ser aplicado tanto numa perspectiva de contornos com região de formas. Para o cálculo deste descritor a forma é projectada numa grelha de tamanho fixo, 16x16 células por exemplo. A cada uma dessas células é atribuído um valor 1 caso contenham uma parte da forma e 0 caso estejam fora da fronteira da forma a descrever. O descritor é a sequência ordenada desses números.

2.4.3 Características de Textura

A textura é um elemento importante na visão humana, que fornece informação acerca da profundidade e orientação numa cena. A análise de texturas 2D tem diversas áreas de aplicação de que são exemplo a inspecção de superfícies ao nível industrial, a análise de imagens biomédicas, e as aplicações de análise via satélite. Em computação gráfica é comum relacionar texturas a superfícies em 3D com o objectivo de aumentar o realismo.

O maior problema relacionado com as texturas do mundo real reside no facto de elas muitas vezes não serem uniformes. Variações na orientação, escala, ou outras variações visuais são factores que contribuem para essa não uniformização. Adicionalmente, Randen & Husøy (1999) referem o elevado grau de complexidade computacional associado ao cálculo das mais comuns medidas de texturas. Num recente trabalho

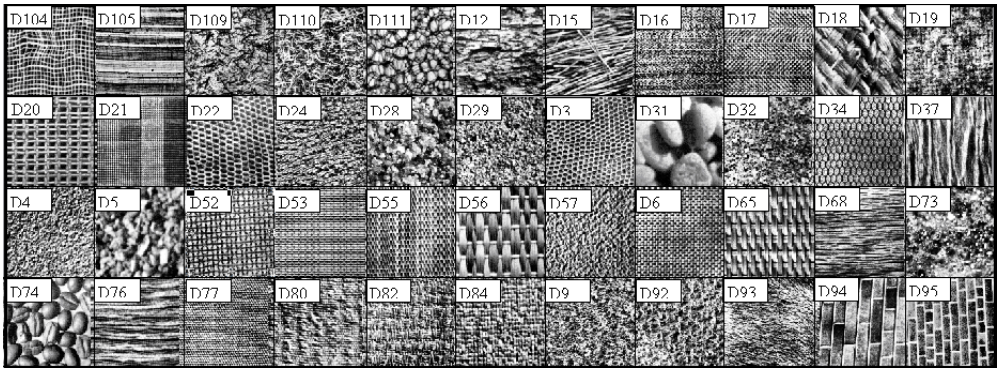


Figura 2.9: Exemplos de texturas. Fonte: Brodatz (1966)

comparativo dos variados métodos de filtragem espacial, Randen & Husøy (1999), dizem que "Os caminhos futuros devem apontar para o desenvolvimento de poderosas medidas de textura que possam ser extraídas e classificadas com baixo poder e complexidade computacional"[3].

Muitas técnicas de classificação de texturas, como forma de simplificação, assumem como constantes a escala, orientação e propriedades de escala de cinzentos. No entanto na classificação de imagens reais essas propriedades não são constantes (podem ocorrer em resoluções espaciais arbitrárias, rotações ou mesmo condicionadas por variações arbitrárias de iluminação), obrigando os algoritmos a várias passagens, fazendo ajustes a essas propriedades em cada passagem (processamento segundo várias orientações: 0º, 45º, 90º, etc.). Isto inspirou o desenvolvimento de algoritmos que incorporem invariância das medidas de textura a pelo menos uma ou duas das propriedades de escala, orientação e escala de cinzentos.

No âmbito da extracção de características invariantes à rotação destacam-se, a extracção de matrizes de co-ocorrência da imagem que captura a distribuição espacial dos contornos em 5 direcções: vertical, horizontal, diagonal a 45º, diagonal a 135º e isotrópicos (Davis *et al.*, 1979), os polarogramas (*Polarograms*) (Davis, 1981), e a anisotropia de texturas (*texture anisotropy*) (Chetverikov, 1982), utilização de filtros no domínio das frequências (ex.: filtros de *Gabor*) (Drimbarean & Whelan, 2001; Manjunath & Ma, 1996), abordagens utilizando filtros MRF (*Markov Random Field*), como são os exemplos do CSAR (*Circular Simultaneous Autoregressive*) proposto por Kashyap & Khotanzad (1986) e do MRSAR (*Multiresolution Simultaneous Autorregressive*) (Mao & Jain, 1992).

[3]traduzido de "*A very useful direction for future research is therefore the development of powerful texture measurements that can be extracted an classified with al low-computational complexity.*"

No caso de abordagens centradas na característica, como é o exemplo da filtragem com *Gabor wavelets* ou outras funções base, a invariância é conseguida através do cálculo de características invariantes à rotação nas imagens filtradas (Fountain & Tan, 1998). Propostas que incorporam invariância à rotação e escala são: HMHV (*Spectrum Estimation and the Fourier-Mellin transform*) (Alata *et al.*, 1998), GMRF (*Gaussian-Markov Random Field*) (Cohen *et al.*, 1991), Classe de Funções Base (Manian & Vásquez, 1998), *Texture Tuned Masks* (You & Cohen, 1993). A abordagem baseada nos Momentos de Zernike (Healey, 1998; Teh & Chin, 1988) foi uma das primeiras a conseguir invariância às três propriedades. Ojala *et al.* (2002c,b) fazem a introdução às LBP (*Local Binary Pattern*).

Descritores MPEG-7 de Textura

- *Edge Histogram* - Este descritor captura a distribuição espacial de contornos, que são agrupados em 5 categorias: vertical, horizontal, 45º diagonal, 135º diagonal e isotrópicos. A imagem original é dividida em sub-imagens 4x4, e para cada uma é calculada a frequência de cada tipo de contorno. Ou seja, de cada sub-imagem resultam 80 (16x5) valores do histograma de contornos. Os valores do histograma da imagem são normalizados para o intervalo [0, 1] e através de uma quantificação não linear é calculada uma representação de 3 bits por valor do histograma. A comparação de histogramas é feita com recurso à normalização L1 (Horn & Johnson, 1990).

- *Homogeneous Texture* - Este descritor é obtido através da filtragem da imagem utilizando um conjunto de 30 filtros de *Gabor*, com 5 orientações para cada cada uma das 6 escalas diferentes. A média e o desvio padrão das imagens filtradas são calculados no domínio das frequências. Também são calculados a média e desvio padrão da imagem original resultando num vector com 62 coeficientes. Posteriormente é aplicada uma escala não linear e uma quantificação em 8 bits para finalização da representação do descritor. A normalização L1 (Horn & Johnson, 1990) também é a forma utilizada para comparação entre descritores.

- *Texture Browsing* - Também faz uso de uma filtragem de *Gabor* a 5 orientações e 6 escalas diferentes. Da análise das imagens filtradas, é calculado um descritor compacto de 12 bits: 2 bits para a regularidade da textura, 3 bits x 2 para a direcção e 2 bits x 2 para a escala. Como as texturas podem ter mais do que uma direcção dominante e escala associada, a especificação admite um máximo de

duas direcções e de dois valores de escala.

- *Local Binary Pattern Descriptor* - O operador LBP (*Local Binary Pattern*) introduzido por Ojala *et al.* (2002c) serve de base de cálculo neste descritor. Através da passagem de um operador LBP previamente escolhido, são acumulados os valores obtidos num histograma. Como os operadores LBP têm um conjunto fixo e discreto de valores de saída não é necessária qualquer normalização ou quantificação da característica. O histograma de frequência é transformado num histograma de probabilidades dividindo cada valor pelo número total de entradas. Cada valor fornece uma estimativa da probabilidade de encontrar o padrão correspondente na imagem. Na fase de recuperação a (di)similaridade de duas imagens é feita através de teste estatístico não paramétrico. Dadas duas imagens Q e D a sua distância de semelhança é dada por:

$$L(Q, D) = \sum_{b=1}^{B} D_b \log Q_b \tag{2.7}$$

onde B é o número de valores e Q_b e D_b correspondem às probabilidades do valor b na duas imagens.

2.4.4 Características de Movimento

As características de movimento em sequências de vídeo fornecem os indicadores de informação temporal necessários à indexação de vídeo. Técnicas de estimação de movimento de câmara, técnicas de identificação de semelhanças em trajectórias e histogramas de vectores agregados de movimento são as técnicas mais utilizadas na indexação de vídeo utilizando características de movimento (Hampapur *et al.*, 1997; Ashley *et al.*, 1995; Chang *et al.*, 1997a).

A descrição de segmentos de vídeo que capturem integralmente a informação de movimento, além de ser computacionalmente complexa, requer o armazenamento de grandes quantidades de informação. A utilização de características de movimento que consigam representar a essência do movimento utilizando representações compactas e concisas é a alternativa à descrição integral de movimento. A extracção de características de movimento no domínio de compressão MPEG-1/MPEG-2 tornou-se popular uma vez que simplesmente retira informação dos vectores de movimento do *bitstream*

(Wikipedia, 2007b) comprimido. As técnicas de extracção de características de movimento de segmentos de vídeo não comprimidos são mais complexos uma vez que requerem poder de computação adicional e análise imagem a imagem.

As características de movimento agrupam-se em dois tipos: movimento em segmentos de vídeo e movimento em regiões (Jeannin & Divakaran, 2001a).

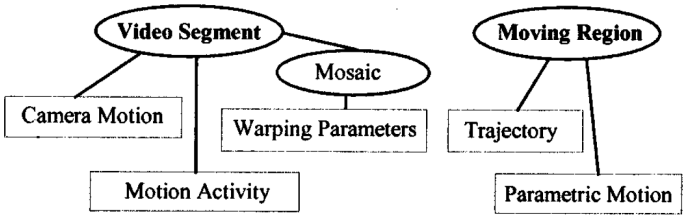


Figura 2.10: Taxonomia dos descritores de movimento. Fonte: Jeannin & Divakaran (2001a)

A actividade geral ou velocidade do segmento é capturada pelas características de Actividade de Movimento (*Motion Activity*). Estas características podem agrupar-se ainda em:

- **Intensidade de Actividade** (*Activity Intensity*) - É expressa por um valor que determina a intensidade de actividade de um determinado segmento de vídeo.
- **Direcção de Actividade** (*Direction of Activity*) - Como um segmento de vídeo pode ter diferentes objectos com valores/direcções de actividade diferentes, esta característica identifica a direcção dominante.
- **Distribuição Espacial de Actividade** (*Spatial Distribution of Activity*) - Esta característica indica se a actividade de movimento está espalhada por várias regiões ou se se encontra delimitada por uma única região. É uma indicação do número e tamanho de regiões com actividade de movimento.
- **Distribuição Temporal de Actividade** (*Temporal Distribution of Activity*) - Expressa a distribuição temporal da actividade de movimento ao longo de um segmento de vídeo.

Os movimentos de câmara ou o ponto de vista da cena são descritos pelas características de Movimento de Câmara (*Camera Motion*). Este descritor expressa que tipos de

movimentos de câmara estão presentes na sequência de vídeo dentro dos tipos de movimentos de câmara possíveis. Também fornece a sua amplitude, localização temporal ou a sua importância em termos de duração (por exemplo: *zoom* em 30% do segmento).

Os movimentos globais ou Mosaicos, como são referidos pela literatura por Szeliski (1994) e Irani *et al.* (1996), e recentemente normalizados pelo MPEG-4 (Koenen, 2002) utilizando a terminologia *Sprite*, são capturados pelos Parâmetros de Trama (*Warping Parameters*). No campo do movimento de regiões incluem-se:

- **Trajectória de Movimento** (*Motion Trajectory*) - Descrevem movimentos de objectos em sequências de imagens;
- **Movimento Paramétrico** (*Parametric Motion*) - Identifica objectos com propriedades de movimento similar.

Marr (1982), descreveu a percepção por humanos de movimentos de trajectórias de objectos como informação de alto nível semântico. Jeannin & Divakaran (2001b) e Ohm *et al.* (1999) através dos chamados *Core Experiments* também verificou que a informação de alto nível recuperada através de descritores de movimento é a suficiente para construir aplicações de recuperação de vídeo de que são exemplos a hiperligação de vídeo (*video hiperlinking*) (Hori & Kaneko, 1999) e as interrogações baseadas em movimento (*motion-based querying*) (Lee *et al.*, 1999).

Descritores MPEG-7 de Movimento

- *Motion Activity* - Este descritor é gerado através do cálculo da matriz de Actividade de Movimento. Esta matriz tem a dimensão da imagem, e cada elemento da matriz reflecte a diferença de cor entre pixels de imagens consecutivas. Quando são observadas diferenças de valores de cor o respectivo elemento da matriz é incrementado. Este processo é repetido para todas as imagens da sequência até chegar à matriz final.
- *Camera Motion* - Este descritor caracteriza informação 3D de movimentos de câmara. Esta informação normalmente é gerada ou capturada directamente por dispositivos de captura de imagem.
- *Motion Trajectory* - Reflecte a informação de trajectórias de movimento de objectos. Este descritor é uma lista de pontos chave *z(x,y,z,t)*. Em conjunto com funções

de interpolação que descrevem percursos de objectos entre pontos chave é possível a representação de informação adicional como por exemplo a aceleração.

- *Parametric Motion* - Baseia-se nos modelos paramétricos (Wikipedia, 2007n). O principio básico é descrever movimentos de objectos em sequência vídeo como sendo um modelo paramétrico 2D.

2.4.5 MPEG-7 eXperimentation Model

Existem diversas propostas e algoritmos de extracção das características descritas nas secções anteriores. A área de extracção de características de imagem e vídeo é uma área de estudo intenso onde a cada passo surgem novas características ou algoritmos mais eficientes e compactos de caracterização audiovisual.

Durante a fase de elaboração da norma MPEG-7 foram definidos vários tipos de descritores e algoritmos associados que iterativamente formaram um pacote de software de referência na área do MPEG-7. O MPEG-7 XM (*eXperimentation Model*) (LIS, 2006) é, assim, uma plataforma de simulação, nascida no seio do processo de normalização, que contém implementação de algoritmos ligados à extracção de características, elaboração de descritores e métodos de similaridade que servem como base de trabalho a desenvolvimento de aplicações MPEG-7. Durante a especificação da norma MPEG-7 todos os componentes dos elementos normativos foram testados, em condições bem definidas, através de contribuições e propostas. Este procedimento deu origem à parte 6 do MPEG-7 que corresponde aos *Core Experiments* (Ohm *et al.*, 1999). Na plataforma do MPEG-7 XM grande parte dos descritores e esquemas de descrição tem pelo menos uma aplicação representativa que elucida a sua funcionalidade e aplicação, tanto na extracção de metainformação, como na anotação de material audiovisual. O MPEG-7 XM foi elaborado de forma modular para simplificar a utilização de componentes isolados ou em conjunto, proporcionado uma boa ferramenta de base para o desenvolvimento de sistemas de recuperação de informação audiovisual. Ao nível da arquitectura, o MPEG-7 XM encontra-se dividido em módulos de 6 tipos que, interligados entre si, formam a chamada cadeia de processamento (Martínez, 2002):

- *Media Decoders* - Este módulo corresponde ao início da cadeia de processamento, é aqui que é carregado o conteúdo audiovisual nos diferentes formatos suportados. O MPEG-7 XM suporta áudio em formato WAV (Wikipedia, 2007u), vídeo

em MPEG-1 e imagens no formato PPM (Netpbm, 2003). Imagens JPEG (Wikipedia, 2007l), GIF (Wikipedia, 2007j) ou PNG (Wikipedia, 2007p) podem ser utilizadas pelo MPEG-7 XM através de uma conversão para o formato PPM com a biblioteca ImageMagick (Still, 2005).

- *Multimedia Data* - Tem a responsabilidade de carregar o material audiovisual para memória e transformar vídeo em sequências de imagens para que possa ser analisado pelos descritores de Cor, Forma, Textura e Movimento.

- *Extraction Tool* - Faz a extracção de características de um único elemento multimédia. Recebe referências para o conteúdo multimédia e respectivo descritor, onde guarda o resultado do processo de extracção.

- *Descriptor Class* - As classes de descrição guardam os dados dos descritores para as partes normativas do MPEG-7. Existem classes para cada um dos Descritores e Esquema de Descrição.

- *Coding Scheme* - O módulo de Esquemas de Codificação contém a implementação da norma para a codificação e descodificação dos Descritores e Esquema de Descrição. A codificação neste âmbito refere-se à forma como o Descritor é escrito e lido de ficheiros. Por exemplo, o MPEG-7 suporta dois tipos de codificação de descritores, uma XML e outra através um formato binário (BiM - *MPEG-7 Binary Format for XML Data*) (Niedermeier *et al.*, 2002) normalizado pelo MPEG-7.

- *Search Tool* - Finalmente este módulo compara descritores MPEG-7 retornando um medida de semelhança entre eles.

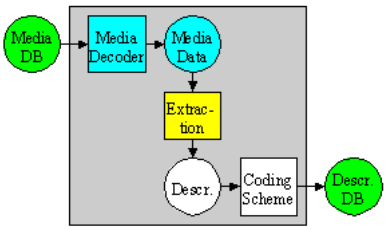


Figura 2.11: Configuração de módulos MPEG-7 XM para aplicação de extracção de descritores. Fonte: Martínez (2002)

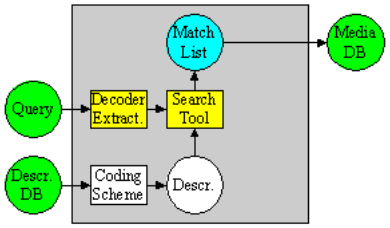


Figura 2.12: Configuração de módulos MPEG-7 XM para aplicação de pesquisa e recuperação. Fonte: Martínez (2002)

2.5 Ontologias

O termo ontologia é oriundo da filosofia onde designa a teoria do ser. Começou a ser utilizado no âmbito da Informática, área de Inteligência Artificial, no início dos anos 90 em projectos cujo objectivo era a organização de grandes bases de conhecimento, tais como CYC (Lenat *et al.*, 1990) e Ontolingua (Gruber, 1992). São várias as áreas que têm aplicado o conceito de ontologia, incluindo a integração inteligente de informação, a cooperação de sistemas de informação, a recuperação de informação e a gestão do conhecimento (Guarino, 1998).

Segundo Gruber (1993): "Uma ontologia é uma especificação formal e explícita de uma *conceptualização* partilhada". Entenda-se *conceptualização* como um modelo abstracto de algum fenómeno no mundo que identifique os conceitos relevantes de tal fenómeno. A especificação deve ser formal, de forma a que seja possível a compreensão por um agente não humano (*software*), e explícita, de forma a que o tipo dos conceitos e as restrições ao seu uso sejam estabelecidos explicitamente.

Acredita-se que a representação formal do conhecimento tenha começado na Índia no primeiro milénio a.C. com o estudo da gramática de *Shastric Sanskrit* (Briggs, 1985). No entanto, da forma com é vista actualmente, a representação formal do conhecimento está muito próxima de trabalhos realizados na Grécia Antiga, principalmente por Aristóteles (384-322 a.C.), nos campos da lógica, das ciências naturais e da filosofia metafísica (Daum & Merten, 2002). O conjunto de estudos realizados nessa área, iniciado por Aristóteles com o seu abrangente sistema de classificação, de taxonomização e de representação do conhecimento de forma geral, é apelidado hoje em dia de ontologia na filosofia.

Tendo em conta o seu significado filosófico, uma ontologia pode ser classificada como

um sistema de categorização que tem por objectivo explicar uma certa visão do mundo, independentemente da linguagem utilizada para descrevê-la. No entanto, uma ontologia no mundo da informática é concretizada como um artefacto de engenharia formado por um vocabulário específico que descreve uma certa realidade (Guarino & Welty, 2000).

A *conceptualização* proposta por Gruber (1993) trata a ontologia filosófica. A *conceptualização* é independente da linguagem na qual é especificada, pelo que que duas ontologias podem ser totalmente diferentes em relação ao vocabulário utilizado, mas referirem-se a uma mesma *conceptualização*.

Os esquemas de bases de dados relacionais podem ser vistos como exemplos de ontologias, pois definem uma hierarquia de tipos, especificando classes e relações de subordinação (Codd, 1990). Ao estabelecer uma formalização de conceitos, uma ontologia permite que haja comunicação com segurança dentro de um contexto ou domínio. Uma ontologia permite a comunicação entre sistemas computorizados independentemente das tecnologias, arquitecturas e domínios (Glushko *et al.*, 1999), e possui um papel crucial no processamento de conhecimento baseado na *web* (Decker *et al.*, 2000).

2.5.1 Ontologia na Web

A elaboração de gramáticas e vocabulários, comuns a uma comunidade de utilizadores num determinado domínio, é um dos resultados do trabalho de definição de ontologias. A uniformização de referências, possibilitando e facilitando o processo de descoberta e geração de conhecimento é assim um dos grandes objectivos das ontologias.

Noy & McGuinness (2001) refere como motivações para a utilização de ontologias a partilha de formas comuns de conhecimento de estruturas de informação entre pessoas e agentes de software, a possibilidade de reutilização de conhecimento de domínio, a explicitação de suposições acerca dos domínios, e a separação de conhecimento de domínio de conhecimento operacional.

No domínio da *web* o *World Wide Web Consortium* (W3C) definiu a *web* semântica (*semantic web*) que pretende ser uma extensão à *World Wide Web*, em que os conteúdos possam ser expressos não só em linguagem natural mas também descritos e interpretados através de agentes de software que possam encontrar, partilhar e integrar informação.

Os documentos da *web* semântica são descritos através de *Resource Description Frame-*

work (RDF) (W3C, 2007b) que promove a integração entre descrições com vocabulários diferentes. Por outro lado, as linguagens *RDF Schema* (RDFS) (W3C, 2007c) e a *Web Ontology Language* (OWL) (W3C, 2007a) são usadas para descrever a semântica utilizada em domínios específicos.

O SchemaWeb (SchemaWeb, 2007) é um repositório de ontologias criado com o propósito de que as várias aplicações e dispositivos possam construir sobre definições já existentes para conceitos.

2.5.2 Classificação de Ontologias

As ontologias podem ser classificadas quanto ao tipo e quanto à profundidade. Guarino & Welty (2000) organizou as ontologias existentes nos seguintes tipos:

- **Ontologias Genéricas** - Não dependem de um problema ou domínio em particular;
- **Ontologias de Domínio** - Aplicam-se exclusivamente a um determinado domínio de conhecimento; uma aplicação exemplo é a área da saúde, que formaliza conceitos como o do médico ou registo do paciente;
- **Ontologias de Tarefas** - Aplicam-se a um determinada tarefa, como exemplo a análise de requisitos de software;
- **Ontologias de Aplicação** - Aplicam-se a um determinada aplicação, especializando Ontologias de Domínio ou Tarefas.

Quanto à profundidade, Guarino & Welty (2000) classificam as ontologias em vários níveis:

- **Nível de Vocabulários** - É a forma mais simples de uma ontologia, podendo ser definida por exemplo através de um XML Schema;
- **Nível de Taxonomia** - A definição de relacionamentos entre os termos estabelece significados, o mais comum dos relacionamentos é o "é um"[4] que formaliza hierarquias do tipo taxonómico;

[4] do inglês *is a*

- **Nível Relacional** - São relacionamentos não hierárquicos, como por exemplo o relacionamento "Orientado por" entre Mestrando e Orientador;
- **Nível Axiomático** - Além de relacionamentos, as Ontologias definem restrições, que são conhecidos como axiomas. Um exemplo de um axioma numa ontologia de Tese de Mestrado seria: "Uma Tese Mestrado deve ter defendida perante um júri".

2.5.3 Tesauro, Dicionário e Vocabulário Controlado

A Wikipedia (2007s) define Tesauro como um "dicionário de ideias afins, (...) uma lista de palavras com significados semelhantes, dentro de um domínio específico de conhecimento". Um Tesauro não deve ser confundido com um dicionário que inclui definições acerca de vocábulos. Também não deve ser visto simplesmente como uma lista de sinónimos. O objectivo do Tesauro é mostrar as relações entre palavras. Um Tesauro aceita relações de sinónimos, antónimo, mais geral (*broader than*), mais específico (*narrower than*).

Uschold & Jasper (1999) apresentam os Tesauros como ontologias simples, uma vez que uma ontologia complexa, segundo estes autores, exige uma riqueza maior nas relações do que as tradicionais apresentadas num Tesauro. Pode citar-se como exemplo a seguinte passagem de Fensel *et al.* (2001): "Grandes Ontologias como a WordNet fornecem um Tesauro para mais de 100,000 termos descritos em linguagem natural"[5]. Neste sentido pode dizer-se que os Tesauros são um tipo de ontologia voltada para a organização de termos. Stojanovic (2005) afirma também que um vocabulário pode ser considerado uma ontologia. O termo Tesauro, que também designa dicionário, vocabulário ou léxico, começou a ser utilizado com mais frequência após a publicação do dicionário analógico de Peter Mark Roget (Wikipedia, 2007o), em Londres, 1852, intitulado "Thesaurus of English Words and Phrases". Diversas definições e significados para o termo surgiram entretanto. Sendo o programa UNISIST (Unesco, 1977) o primeiro a analisar o termo sob aspecto estrutural e funcional, definiu estruturalmente Tesauro como "um vocabulário controlado dinâmico de termos relacionados semântica e genericamente cobrindo um domínio específico do conhecimento", e funcionalmente como "um dispositivo de controle terminológico usado na tradução da linguagem

[5] "*Large Ontologies such as WordNet provide a thesaurus for over 100,000 terms explained in natural language*".

natural dos documentos, dos indexadores ou dos utilizadores numa linguagem do sistema (linguagem de documentação, linguagem de informação) mais restrita".

Da riqueza das ontologias e da disponibilização da linguagem OWL de definição e instanciação de ontologias na *Web* é possível o tratamento automático do equivalente aos Tesauros. Esta tecnologia ainda é emergente, não existindo ainda nenhuma ferramenta disponível que tire partido do OWL.

2.6 Interacção com Utilizador

Há três formas clássicas de pesquisar material multimédia previamente arquivado:

- **Navegação** - Navegar (*browse*) sobre uma colecção de imagens, áudio e ficheiros vídeo, até ser encontrado o que se procura;
- **Pesquisa Baseada em Texto** - Informação textual (metainformação) é adicionada ao conteúdo audiovisual durante o processo de catalogação. Na fase de recuperação, esta informação é utilizada como base de pesquisa utilizando as técnicas convencionais de pesquisa sobre texto para encontrar o conteúdo audiovisual associado;
- **Pesquisa Baseada em Conteúdo** - Pesquisa sobre repositórios multimédia utilizando informação acerca do conteúdo da imagem, áudio ou vídeo. Esta informação de conteúdo é mapeada utilizando técnicas adequadas que consigam identificar semelhanças em outros conteúdos audiovisuais.

Tanto a pesquisa livre como a baseada em texto tem algumas limitações e problemas de escalabilidade. A pesquisa livre é um bom método para uso esporádico ou quando o utilizador não sabe bem ao certo o que pretende encontrar, ou seja, para utilizadores que necessitam de obter resultados com rapidez e se um grau de detalhe elevado não é viável. Além disso é um método moroso, que requer grande paciência, e ineficiente, o que o torna completamente inadequado para bases de dados grandes.

A pesquisa baseada em texto tem dois problemas associados. O primeiro é a necessidade de despender tempo significativo para anotar manualmente cada imagem ou cena. O segundo é a imprecisão associada à percepção humana acerca dos conteúdos,

da qual resultam diferenças no detalhe e exaustividade com que os conteúdos são anotados por pessoas diferentes, e uma grande dependência da qualidade da anotação em relação aos conhecimentos do individuo que faz a anotação.

A pesquisa baseada no conteúdo acrescenta um valor adicional à pesquisa baseada em texto, uma vez que que podem ser utilizadas técnicas do domínio da visão por computador para encontrar semelhanças entre conteúdo visual.

No que respeita às técnicas de interacção com o utilizador podem ser identificados os seguintes tipos de pesquisa (Marques & Furht, 2002):

- **Pesquisa Interactiva** - É um tipo de pesquisa ideal quando o utilizador não tem ideia formada acerca do que pretende encontrar. Podem ser utilizadas técnicas de agrupamento (*clustering*) de imagem para agrupar e organizar visualmente imagens similares. Desta forma minimiza-se o número de imagens não pretendidas pelo utilizador.

- **Navegação por Categoria** - Pesquisa por categorias ou assuntos, num conjunto ou subconjunto de hierarquias, até se conseguir reduzir o número de imagens a pesquisar.

- **Pesquisa por Imagem Exemplo** - Através da especificação de uma imagem o utilizador obtém uma lista de imagens semelhantes ordenadas por ordem decrescente de relevância.

- **Pesquisa por Esboço ou Desenho** - O utilizador especifica um esboço acerca da imagem ou vídeo que tem em mente pesquisar.

- **Pesquisa por Características Visuais** - Pesquisa directa de características visuais (ex.: cor, textura, forma, propriedades de movimento). É mais indicada para utilizadores que já têm um conhecimento mais concreto e precisam de efectuar pesquisas detalhadas.

- **Pesquisa por Palavra Chave ou Texto** - Pesquisa sobre imagem ou vídeo que tem associadas palavras-chave introduzidas anteriormente ou texto retirado de diálogos ou legendas em imagens ou vídeo.

2.6.1 Fosso Sensorial e Fosso Semântico

Dependendo do modelo de interrogação utilizado para pesquisas em sistemas de recuperação audiovisual, os domínios de interrogação e de documentos podem ser diferentes.

Um exemplo desta diferença é a utilização de interrogações de texto num conjunto de imagens com o objectivo de recuperar imagens relevantes para a interrogação textual. Mesmo quando se tem o mesmo domínio de interrogações e documentos pode haver dificuldade em fazer coincidir conceitos envolvidos nessas interrogações. Há portanto diferenças entre o resultado esperado da interrogação introduzida por um utilizador e o resultado obtido.

Num contexto de recuperação de informação visual baseada em conteúdo essa diferença é justificada por duas razões: o fosso sensorial (*sensory gap*) e o fosso semântico (*semantic gap*).

O fosso sensorial diz respeito a diferenças entre objectos observados segundo condições diferentes, ou seja, duas imagens de um mesmo objecto podem ser totalmente diferentes devido a diferenças nas condições de luz, escala, rotação. Os problemas relacionados com a representação de objectos também são do domínio sensorial e incluem diferenças no tipo de codificação, diferenças no níveis de quantificação e esquemas de compressão (com perdas ou sem perdas), podendo dificultar a identificação de semelhanças entre objectos.

Uma definição para fosso sensorial é a dada por Gevers & Smeulders (2004): "O fosso sensorial é a diferença entre o objecto no mundo real e a informação numa descrição (computacional) de uma gravação dessa cena"[6].

A recuperação de informação baseada em conteúdo é assente em múltiplas características de baixo nível (cor, forma, textura, movimento) para descrever conteúdos audiovisuais. Para conseguir lidar com o fosso sensorial, essas características devem ser consistentes e invariantes por forma a conseguir manter-se representativas do conjunto de objectos audiovisuais. Num contexto de recuperação de imagem, um utilizador pode servir-se de uma imagem para recuperar imagens semelhantes, ou seja, imagens cujo conjunto de características de baixo nível seja semelhante. No entanto ao nível

[6]do inglês "*The sensory gap is the gap between the object in the world and the information in a (computational) description derived from a recording of that scene.*"

semântico isso já não é verdade, uma vez que imagens com características de baixo nível semelhantes podem corresponder a imagens totalmente distintas do ponto de vista semântico.

Segundo Gevers & Smeulders (2004) "o fosso semântico é a distância existente entre a informação que pode ser extraída de dados visuais e a interpretação que esses mesmos dados tem por parte de um utilizador numa determinada situação"[7].

Um utilizador pretende fazer pesquisas em imagens a um nível conceptual, ou seja, recuperar imagens que contenham objectos de um determinado tipo ou que sejam relativas a um determinado tema ou categoria. As descrições automáticas das imagens são derivadas de características de baixo nível (sem informação semântica ou de contexto), revelando os problemas do fosso semântico. A total associação de características de baixo nível com a informação semântica que elas representam é no fundo a solução para o fosso semântico. É um problema que não tem ainda solução, estando a investigação neste domínio centrada na atenuação desse problema utilizando técnicas auxiliares.

A adição de etiquetas, palavras-chave ou anotações de texto é a forma mais usual de atenuar o problema do fosso semântico, mas esta abordagem é um processo manual e moroso. Outras técnicas são as propostas por Chang *et al.* (1997b) e Santini *et al.* (1999), que utilizam programas que exploram a *Web* coleccionando imagens e inserindo-as numa taxonomia predefinida partindo de informação relacionada que se encontra perto de elas. Uma proposta similar nesta área é também a de Chen *et al.* (1996) que transpõem essa técnica para as bibliotecas digitais.

2.6.2 Realimentação de Relevância

Como já descrito anteriormente a resolução ou atenuação do fosso semântico é um dos objectivos de qualquer sistema de recuperação de conteúdo multimédia. Marques & Furht (2002) identificaram duas formas de atenuar esse problema, sendo a primeira baseada na adição da maior quantidade possível de metainformação aos conteúdos multimédia. Essa metainformação pode ser inserida por humanos, o que implica grandes esforços de anotação ou catalogação de conteúdos. Em alternativa podem usar-se técnicas de adição de metainformação automáticas também descritas na secção anteri-

[7] do inglês "*The semantic gap is the lack of coincidence between the information that one can extract from the visual data and the interpretation that the same data have for a user in a given situation.*"

or. A outra forma identificada por Marques & Furht (2002) é a utilização de técnicas de interacção com o utilizador como forma de conseguir que o sistema identifique o contexto semântico dos conteúdos. A utilização de técnicas de realimentação de relevância (*Relevance Feedback*) em conjunto com algoritmos de aprendizagem pode fornecer ao sistema informação semântica útil no processo de recuperação.

O conceito por trás da realimentação de relevância é tomar os resultados que são inicialmente retornados, após uma interrogação, e usar informação submetida pelo utilizador para saber se esses resultados são relevantes para o utilizador ou não. Em suma, a realimentação de relevância é o processo pelo qual um sistema obtém informação dos utilizadores acerca da relevância de características, imagens, regiões de imagens, ou resultados parciais obtidos. Como descrito por Gevers & Smeulders (1999), o objectivo da realimentação de relevância é o de envolver o utilizador no processo de formulação de interrogações através da especificação de conteúdos relevantes ou não relevantes.

Dependendo da sua utilização a realimentação de relevância caracteriza-se segundo três tipos:

- **Feedback Explícito** - É obtido partindo da especificação por parte do utilizador dos documentos relevantes e não relevantes;

- **Feedback Implícito** - É inferido através do comportamento do utilizador, memorizando quais os documentos escolhidos pelo utilizador para visualização e os não escolhidos;

- **Feedback Cego** - É obtido partindo do pressuposto de que os n primeiros documentos da lista de ordenação são relevantes.

A utilização da informação de realimentação de relevância pode ser feita através da formulação de novas interrogações, ajustando os pesos dos termos da interrogação original, incluindo essa informação na lista de termos da interrogação. A realimentação de relevância é normalmente implementada usando o algoritmo de Rocchio (Joachims, 1997).

2.7 Sistemas Existentes

Os sistemas de recuperação de imagem e vídeo existentes, onde as interrogações podem ser feitas de forma visual, agrupam-se segundo dois tipos:

- **Interrogação segundo um Exemplo** (*Query by Example*) - A pesquisa é efectuada através da de uma imagem exemplo (fornecida pelo utilizador ou escolhida de um determinado conjunto). O sistema pesquisa imagens similares baseada em critérios de semelhança de características de baixo nível.

- **Esboços Visuais** (*Visual Sketches*) - Através da especificação, por parte do utilizador, de um esboço ou desenho aproximado de uma imagem (blocos de cor, formas, padrões de texturas, etc.), o sistema pesquisa imagens que coincidam morfologicamente com o esboço introduzido através de comparação de características de baixo nível.

Sistemas na área da recuperação de imagem que utilizam essas técnicas são: o Query by Image Content (QBIC) da IBM (Ashley *et al.*, 1995), o VisualSEEk (Smith & Chang, 1996), o Photobook (Pentland *et al.*, 1994), o blobworld (Carson *et al.*, 1999) e o Retrievr (Jacobs *et al.*, 1995). O Virage Video Engine (Hampapur *et al.*, 1997), o CueVideo (Ponceleon *et al.*, 1998) e o VideoQ (Chang *et al.*, 1997a) são exemplos de sistemas de recuperação na área do vídeo.

No domínio da *Web* os conceitos de pesquisa de imagens por similaridade de características de baixo nível também são utilizados no Webseek (Smith & Chang, 1997) e no Webseer (Frankel *et al.*, 1996).

2.8 Avaliação de Sistemas de Recuperação de Informação

A eficiência e eficácia são dois factores fundamentais na avaliação de um sistema de recuperação. No processo de avaliação de um sistema de recuperação de informação devem ser considerados vários factores: o tempo gasto para realização da pesquisa; o tempo dispendido pelo utilizador para obter a informação desejada; e a capacidade do sistema para recuperar documentos realmente relevantes.

Um método de avaliação muito divulgado é a comparação de sistemas utilizando colecções de documentos de referência. Essas colecções de documentos de teste contêm um conjunto de interrogações e a avaliações de relevância de cada documento para cada interrogação, resultando numa boa medida comparativa entre sistemas (Zobel, 1998). O TREC (*Text Retrieval Conference*) (NIST, 1992) é uma iniciativa de avaliação de colecções de testes que produz resultados no âmbito de *workshops* na área de recuperação de informação (Harman, 1993). Na área multimédia o TRECVID (*TREC Video Retrieval Evaluation*) (NIST, 2003; Smeaton *et al.*, 2004) fornece resultados de avaliações de colecções de testes de imagem e vídeo.

Tendo como referencia os dados divulgados no TREC e TRECVID é possível comparar os resultados de um sistema de recuperação de informação produzidos pelos grupos que participam no TREC ou TRECVID, conseguindo-se, assim, obter valores de recuperação[8] (*recall*) e precisão (*precision*) para o sistema que está a ser avaliado.

A medida recuperação indica a capacidade do sistema de recuperação de informação de recuperar os documentos relevantes para a consulta em questão. Quanto maior for o número de documentos relevantes recuperados, maior será o valor de recuperação. O seu valor é dado pelo rácio entre o número de documentos relevantes recuperados e o número total de documentos relevantes existentes na colecção de documentos.

$$recuperacão = \frac{\mid DocumentosRelevantesRecuperados \mid}{\mid TotalDocumentosRelevantes \mid} \tag{2.8}$$

A medida de precisão indica a capacidade de um sistema recuperar apenas o que é relevante, ou seja a capacidade do sistema em não recuperar documentos não relevantes para a interrogação. Quanto menor for o número de documentos irrelevantes apresentados, maior será a precisão do sistema.

$$precisão = \frac{\mid DocumentosRelevantesRecuperados \mid}{\mid DocumentosRecuperados \mid} \tag{2.9}$$

Normalmente existe uma relação inversa entre os valores de recuperação e precisão: quando o valor de recuperação é alto, o valor de precisão tende a diminuir e vice-versa, em valores elevados de precisão a recuperação tende a diminuir. Isto deve-se ao facto de que para atingir valores de recuperação elevados, o sistema tende a recuperar

[8] alguns autores utilizam também o termo revocação

um maior número de documentos (relevantes e não relevantes) o que por sua vez faz com que aumente o número de documentos irrelevantes e consequentemente baixe o valor de precisão. Para atingir valores de precisão elevada, o número de documentos recuperados é reduzido, fazendo com que o número de documentos relevantes recuperados possa ser baixo comparado com o total de documentos relevantes existentes (diminuição do recuperação) (Manning & Schtze, 1999; Witten *et al.*, 1999). Normalmente os sistemas são desenhados atendendo a um equilíbrio entre os dois valores.

Outra medida utilizada é a precisão aplicada a um determinado valor de corte (*cutoff*)[9], que não é mais que um ponto de corte entre os documentos recuperados, ou seja, a medida de precisão é realizada apenas nos *n*-primeiros documentos recuperados. Este valor pode indicar a eficácia de um sistema em estabelecer uma ordenação de documentos, fazendo com que os mais relevantes apareçam no inicio da lista (Manning & Schtze, 1999).

[9] sinónimo de valor de limiar (*threshold*)

Capítulo 3

Reutilização de Anotações

A anotação de material multimédia pode ser feita através de processos manuais ou, mais recentemente, de forma automática, ou semi-automática.

A anotação automática é comummente baseada em extractores de características de baixo nível, que coleccionam informação de baixo valor semântico e o transformam em anotações do material audiovisual. A conjugação com descrições textuais retiradas de legendas ou transcrições das pistas de áudio e a utilização de informação de contexto acerca do tipo e assunto do material multimédia ajudam na obtenção de informação de mais alto nível semântico. A extracção automática de informação de alto valor semântico requer a utilização de extractores de características específicos para o domínio de aplicação e tipo de conteúdo. Esse facto torna o processo de extracção de características muito específico e direccionado para um determinado domínio, resultando num fraco desempenho quando aplicado de forma genérica.

Por outro lado, a anotação manual é um processo dispendioso, devido ao facto de envolver trabalho humano. É uma tarefa morosa, porque para anotar vídeo, em média, é necessário um tempo superior em 10 vezes à sua duração (estima-se que 1 minuto de vídeo demore aproximadamente 10 minutos para anotar). A repetição de cenas similares num mesmo segmento ou entre segmentos vídeo leva a repetições no trabalho de anotação. O nível de conhecimento do indivíduo que faz anotação influi na completude, quantidade e qualidade das anotações produzidas. Mesmo assim, actualmente as anotações de qualidade são apenas conseguidas através de anotação manual.

À medida que cresce a quantidade de material multimédia e respectivas necessida-

des para pesquisa e recuperação, a informação de grande qualidade obtida através de anotação manual só poderá continuar a ser uma alternativa viável se o seu custo, tempo de anotação e uniformização for optimizado. A utilização de interfaces visuais melhoradas, que permitam optimizar o processo de anotação manual, a utilização de anotadores com bons conhecimentos de domínio (anotador especializado em desporto automóvel, especialista em política internacional), são factores que podem reduzir o custo da anotação por unidade de tempo.

Existem actualmente diversas ferramentas para anotação de vídeo, podendo ser elas colaborativas ou não. A secção seguinte dá o exemplo de algumas destas ferramentas. A estruturação de informação durante o processo de anotação é também um aspecto importante no tipo de utilização alvo para a aplicação; a Secção 3.2 enumera algumas das normas usadas. Na Secção 3.3 é descrita a mudança de paradigma de anotação, seguindo-se a proposta de um sistema de reutilização de anotação, na Secção 3.4. Para finalizar o capítulo a Secção 3.5 aborda os cenários de utilização onde se espera que a proposta traga vantagens claras.

3.1 Ferramentas de Anotação

Existem actualmente diversos sistemas de indexação e anotação de vídeo no formato digital. Grande parte desses sistemas são de utilização autónoma (*stand-alone*) ou estão inseridos em sistemas de produção audiovisual cujo objectivo é a anotação de material audiovisual por um único utilizador. Ultimamente assistiu-se também ao surgimento de aplicações de anotação vídeo que tiram partido da internet de banda larga, disponibilizando conteúdos vídeo e formas colaborativas de anotação.

Estes dois tipos de sistemas são distintos na forma como o processo de anotação é executado (ferramentas não colaborativas/colaborativas), e também no que respeita à utilização de esquemas de dados com vocabulário controlado. Enquanto nos sistemas de anotação não colaborativos é comum a utilização de esquemas de dados com vocabulário controlado (esquemas de dados normalizados ou específicos ao domínio de aplicação), os sistemas de anotação colaborativa não utilizam qualquer tipo de esquema de dados para anotação. A utilização de etiquetas (*tags* ou *labels*) é bastante comum neste tipo de sistemas. Recentemente tecnologias mais leves como as folksonomias (*folksonomies*) propostas por Mathes (2004) captam esta abordagem no enriquecimen-

to e pesquisa relativa a documentos não estruturados. Os sistemas de recuperação de imagem e vídeo são dois domínios de aplicação onde conceitos de alto nível podem ser obtidos com recurso a etiquetas introduzidas no âmbito de ambientes colaborativos. Exemplos de sistemas deste tipo, na área de recuperação de imagem são o Flickr.com (flickr, 2006; Wikipedia, 2006b) e o del.icio.us (del.icio.us, 2006; Wikipedia, 2006a) na área de recuperação de páginas na *web*. Assistimos portanto a uma nova forma de anotação de conteúdos que é colaborativa e de custo reduzido uma vez que é feita directamente na *web*. Os sistemas de anotação colaborativa de vídeo herdaram as folksonomias dos sistemas de anotação de imagem e organização de páginas, sendo o YouTube (Hurley *et al.*, 2006; Wikipedia, 2007w) um exemplo de crescente utilização de folksonomias em sistemas de recuperação vídeo.

A utilização de técnicas de análise de conteúdo como forma de automatização do processo de anotação é ainda exclusiva dos sistemas de anotação de utilização autónoma e não colaborativa, sendo a sua maior incidência na análise e detecção de cortes de cena.

A anotação de vídeo colaborativa pode ser ainda realizada de forma síncrona ou assíncrona. Quando vários utilizadores num mesmo instante efectuam uma anotação em conjunto, por exemplo numa sala virtual de anotação, diz-se que o processo de anotação é síncrono. No caso de o processo de anotação colaborativa ser paralelo e independente diz-se que é assíncrono. Exemplos de sistemas dos dois tipos são apresentados a seguir.

3.1.1 Anotação Vídeo Não Colaborativa

- ***IBM - MPEG-7 Annotation Tool*** - Este sistema, proposto por Smith & Lugeon (2000), faz detecção de cortes de cena e permite a utilização de um esquema de dados configurável seguindo a norma MPEG-7. Permite a adição de palavras-chave de conteúdo livre e fornece ao utilizador a possibilidade de fazer anotações utilizando um esquema de dados controlado ou a inserção de etiquetas. O sistema tira partido de informação de baixo nível na análise de vídeo para detecção de cortes.

- ***Ricoh - Movie Tool*** - A ferramenta Movie Tool (ricoh, 2005) utiliza uma linha temporal (*timeline*) para representar as anotações num segmento vídeo, usa o MPEG-7 como norma de anotação e fornece detecção de cortes de cena como auxílio à

anotação. Não utiliza qualquer tipo de análise vídeo para geração de anotações automáticas.

- **ZGDV - Videto** - Apesar de abstrair o esquema de dados utilizado, o *Videto* (ZGDV, 2005) faz anotação de vídeo baseado na norma MPEG-7.

- ***Coala - LogCreator*** - O *LogCreator* (EPFL, 2005) é uma ferramenta *web* que suporta descrição de vídeo. Fornece detecção de cortes de cena e utiliza o MPEG-7 como norma de anotação. É uma ferramenta específica para a anotação de documentos de notícias de televisão utilizando uma estrutura de dados pré-definida.

- ***CSIRO's CMWeb Tools*** - O *CMWeb Tools* (CSIRO, 2005) utiliza um formato HTML privativo para geração de páginas *web* com descrições dos vídeos.

- ***iVas*** - O sistema *iVas* (Yamamoto & Nagao, 2004) pode associar várias anotações textuais a segmentos vídeo. O sistema faz análise de vídeo para detecção de cortes e geração de informação de histogramas de cor.

- ***VideoAL*** - O *VideoAL* (Lin *et al.*, 2003) faz uso de informação de conteúdo e algoritmos de aprendizagem para efectuar etiquetagem automática de segmentos vídeo. Não utiliza nenhum esquema de dados normalizado.

- ***theScribe*** - Este sistema (MOG-Solutions, 2007) é baseado em MXF (Wells *et al.*, 2006; SMPTE, 2004b) e utiliza a norma DMS-1 (SMPTE, 2004a) para anotação de conteúdos. Ao nível de análise de conteúdos apenas efectua detecção de cortes de cena.

3.1.2 Anotação Vídeo Colaborativa

- ***YouTube*** - O *YouTube* (Hurley *et al.*, 2006; Wikipedia, 2007w) tornou-se muito popular nos últimos anos e não faz uso de qualquer esquema de dados ou vocabulário controlado nem usa análise vídeo para geração de anotações. Os utilizadores aplicam etiquetas como forma de anotação. O processo de anotação é assíncrono.

- ***Microsoft's MRAS*** - Bargeron *et al.* (2001) propuseram uma aplicação baseada na *web* cujo objectivo é fornecer uma ferramenta aos estudantes para fazerem e partilharem anotações de vídeos das aulas. O processo de anotação é assíncrono.

- ***SMAT*** - O sistema *SMAT* (Steves *et al.*, 2001) permite aos utilizadores adicionarem colaborativamente anotações a conteúdos multimédia através de texto e um quadro virtual (*whiteboard*). Não utiliza qualquer esquema de dados e o processo de anotação é síncrono.

- ***eSport*** - O sistema *eSport* (Zhai *et al.*, 2005) é uma ferramenta de anotação de vídeo colaborativa e síncrona. É um sistema que se encontra dimensionado para suportar anotação de conteúdos desportivos através da *Web* e de um ambiente multi-plataforma.

3.2 Normas de Metainformação

Na secção anterior foram já mencionadas algumas formas de anotação de conteúdos. A forma mais simples de efectuar anotações é sem dúvida o texto livre. Facilmente podemos efectuar anotações de conteúdos, imagens ou até mesmo texto, com recurso a descrições de texto. Esta abordagem, que do ponto de vista de anotação é a mais simples, revela alguns problemas do ponto de vista de recuperação. Primeiro, ao não seguir qualquer esquema de normalização torna a interpretação de dados na fase de recuperação mais complexa. Segundo, torna difícil a sua organização e catalogação. A anotação textual é também demasiado subjectiva sendo difícil garantir a consistência das anotações. O tratamento automático de dados (importação/exportação para outros sistemas) é complexo.

A utilização de taxonomias de termos pré-definidos, como a *WordNet* (Fensel *et al.*, 2001) que estabelece relações semânticas entre termos, pode ajudar a resolver alguns desses problemas. A popularidade de alguns sistemas de anotação por etiquetas reside na minimização dos pontos negativos da anotação textual. Um exemplo do uso de etiquetas são as já referidas folksonomias propostas por Mathes (2004). Todas as etiquetas têm o mesmo peso e a recuperação pode ser efectuada com recurso à utilização de palavras-chave. Ao nível de organização uma abordagem comum é o agrupamento de conteúdos com as mesmas etiquetas (ou eventualmente sinónimos) construindo mapas de ordenação usando para tal o número total de ocorrências de cada uma das etiquetas. Assim etiquetas que apareçam um maior número de vezes terão mais importância.

Os dados introduzidos nas descrições textuais são geralmente anotações descritivas.

Dos processos de análise automática de conteúdos são obtidos valores que caracterizam cor, textura, forma entre outras. Surge então a necessidade de integrar e normalizar estes dois mundos distintos de anotação: descrições (conceitos de alto nível semântico) e características (conceitos de baixo nível semântico).

A normalização e estruturação de esquemas de dados traz vantagens claras na representação de informação e é uma área activa de investigação no domínio dos conteúdos multimédia.

A descrição de um conteúdo multimédia pode ser visto segundo várias perspectivas, dependendo por exemplo do ponto de vista de quem produz ou utiliza a informação:

- **Do ponto de vista do gestor de conteúdos** - As descrições introduzidas são do tipo bibliográfico: autor, título, data de criação, formato de codificação, entre outros. Esta metainformação é voltada para a gestão de conteúdos;
- **Do ponto de vista do fornecedor de serviços** - A descrição adicionada é normalmente de conteúdo e direccionada para a recuperação. Nesse tipo de descrição encontram-se descrições de disponibilidade de formatos, disponibilidade de fontes, e informação semântica: catalogação por género de conteúdo (desporto, lazer, humor). No geral essas descrições servem para melhorar o processo pesquisa em aplicações multimédia;
- **Do ponto de vista do consumidor de conteúdos** - Descrições que incluam as preferências pessoais e disponibilidade de conteúdos ajudam a personalizar a pesquisa de conteúdos.

Ainda dentro dos esquemas de normalização de dados, há que considerar dois tipos de esquemas. Os primeiros permitem apenas a estruturação de metainformação do tipo descritiva ou de estrutura: *Dublin Core* (Weibel & Lagoze, 1997; DCMI, 2000), *NewsML* (Fahy *et al.*, 2003; IPTC, 2000), VAML (Zhou & Jin, 2004), *TV-Anytime* (TV-Anytime, 2003). Os segundos, para além da normalização de esquemas de dados para metainformação descritiva e estrutural, consideram também a forma de estruturação de características de conteúdo (cor, textura, movimento, forma) por forma a facilitar a recuperação baseada em conteúdo: MPEG-7 (Martínez, 2002; Manjunath, 2002; Nack & Lindsay, 1999a,b). Outros formatos têm ainda preocupações de estruturar metainformação relativa à produção, consumo e interoperabilidade entre conteúdos multimédia: MPEG-21 (Bormans & Hill, 2002; Burnett *et al.*, 2006) e DMS-1 (SMPTE, 2004a). De

referir que a norma mais utilizada no âmbito da anotação/recuperação de informação baseada em conteúdo é o MPEG-7 (Manjunath, 2002). A normalização de diversos descritores de baixo nível para mapeamento de informação de conteúdo facilita a uniformização, comparação e interoperabilidade entre sistemas baseados em características de baixo nível.

A configuração ou extensibilidade das normas de estruturação por forma a acompanhar futuras evoluções tecnológicas é também uma característica a avaliar. Neste ponto as normalizações baseadas em XML respondem a esses requisitos.

A importação/exportação e conversão entre diferentes normas pode, nalguns casos, ser efectuado com recurso a operações de transformação o que em geral, e dada a cobertura diferente das várias normas, pode não ser tarefa fácil. A iniciativa *Metadata Crosswalks* (Godby *et al.*, 2004) aborda o problema de correspondência entre conceitos de diferentes normas.

Nas subsecções seguintes são abordadas algumas das normas de estruturação de metainformação sendo a separação feita entre aquelas que suportam anotação genérica de conteúdo e as que foram desenvolvidas num âmbito de um domínio específico.

3.2.1 Normas Genéricas para Anotação de Conteúdo

Dublin Core

O *Dublin Core* (Weibel & Lagoze, 1997; DCMI, 2000) é uma norma para descrição que contém elementos de metainformação com o objectivo de ajudar à descoberta de recursos electrónicos. Estes recursos electrónicos são além de texto (utilização mais vulgar do *Dublin Core*), documentos não textuais como imagem, áudio e vídeo. A norma focou-se em fornecer extensões aos elementos nucleares (*core elements*) através da especificação de sub-elementos e esquemas específicos a dados audiovisuais. Elementos nucleares simples são título (*title*), criador (*creator*), assunto (*subject*), descrição (*description*), editor (*publisher*), contribuidor (*contributor*), data (*date*), tipo (*type*), formato (*format*), identificador (*identifier*), fonte (*source*), língua (*language*), relação (*relation*), âmbito (*coverage*) e direitos de autor (*rights*). O *Dublin Core* é actualmente utilizado como norma de metainformação em muitos arquivos de televisão (Parmar, 2005). O CPB (2005) é um exemplo de um arquivo baseado no *Dublin Core*, enquanto o Zope (2005), Plone (2000) e o Nuxeo (2000) são gestores de conteúdos que o utilizam.

MPEG-7

As normas MPEG-1 (Chiariglione, 1996), MPEG-2 (Chiariglione, 2000) e MPEG-4 (Koenen, 2002) têm como objectivo a representação codificada de informação audiovisual. A norma MPEG-7 (Martínez, 2002) surgiu com o intuito de normalizar uma interface para descrição de material multimédia, ou seja a representação de informação sobre os conteúdos mas não os conteúdos em si. A este conceito dá-se o nome de metainformação.

A especificação de descritores de baixo nível é apenas uma das componentes da norma MPEG-7. A integração de anotações provenientes de diversas fontes num único conjunto de estruturas de dados foi também um dos objectivos da norma. Aspectos como interoperabilidade e globalização de fontes de dados, bem como a flexibilização da sua gestão são objectivos na norma MPEG-7. A uniformização de conceitos das comunidades das bases de dados e processamento de sinal levaram à elaboração de uma norma que satisfizesse ambos os mundos.

Os conceitos base do MPEG-7 (Martínez, 2002) são os dos Esquemas de Descrição (*Description Schemes*), Descritores (*Descriptors*) e Linguagem de Definição de Descritores (*Description Definition Language*).

Os Esquemas de Descrição são estruturas de metainformação escritas em XML que descrevem a estrutura e semântica de conteúdos multimédia. Foram concebidos para representar conceitos de alto nível semântico como por exemplo regiões em imagem ou texto. Os Descritores por sua vez representam características ou descrições de conteúdos. A norma MPEG-7 é geralmente referida devido aos descritores de baixo nível semântico como cor, textura ou forma, uma vez que estes não são suportados por outras normas.

A conjugação de conjuntos de Esquemas de Descrição e Descritores que satisfaçam as necessidades relativas à criação e utilização de metainformação serve a grande maioria das aplicações multimédia. No entanto, o MPEG-7 é extensível através da Linguagem de Definição de Descritores que permite a criação ou modificação dos Esquemas de Descrição ou Descritores para os restantes casos.

No caso das estruturas de dados concebidas para a descrição e anotação de conteúdos audiovisuais, foram criados os Esquemas de Descrição Multimédia (*Multimedia Description Tool*), que são compostos por um conjunto de Esquemas de Descrição e

Descritores.

MPEG-21

Como complemento à normalização de descrições multimédia proposta pelo MPEG-7, o MPEG-21 centra-se nos aspectos de organização e infra-estrutura de sistemas multimédia distribuídos onde é necessária mais do que a informação de objectos individuais.

A norma MPEG-21 propõe o conceito de entidade de distribuição e validação: o Item Digital (*Digital Item*). Este Item Digital é utilizado na interacção com todos os *actores*[1] num sistema multimédia distribuído. A gestão de conteúdos, a gestão de propriedade intelectual e a adaptação de conteúdos é nesta norma estabelecida de forma a suportar diferentes tipos de serviços.

3.2.2 Normas para Domínios Específicos

Dicionário de Metainformação

O SMPTE normalizou um dicionário de metainformação do SMPTE (*SMPTE Metadata Dictionary*) (SMPTE, 2007). O dicionário é uma colecção registada de nomes e tipos de dados, desenvolvido pelos membros da indústria de televisão e vídeo que formam o SMPTE. É uma estrutura hierárquica que permite extensões e mecanismos para formatação de dados em sinais de vídeo e televisão. Além disso fornece também um método comum para a sua implementação. Grande parte da metainformação está sob a forma de atributos específicos dos conteúdos, como por exemplo informação temporal. A anotação semântica não é ainda possível mas existe já uma norma para anotação descritiva de conteúdos o DMS-1 (*Descriptive Metadata Scheme*) (SMPTE, 2004a).

NewsML

A norma NewsML (Fahy *et al.*, 2003; IPTC, 2000), proposta pelo IPTC (*International Press Telecommunications Council*) (IPTC, 1965), é uma norma baseada em XML que visa

[1] No MPEG-21 é dado o nome de *actor* aos utilizadores.

a estruturação de notícias em formato multimédia (texto, áudio, vídeo). Foi concebida para fornecer uma ferramenta (*framework*) de estruturação de notícias independente do formato, permitindo aos editores trabalhar de forma rápida e eficiente num ambiente de notícias totalmente digital. Esta norma beneficia por consequência os consumidores de conteúdos.

SportsML

A norma SportsML (IPTC, 2001), também da responsabilidade do IPTC (*International Press Telecomunications Council*) (IPTC, 1965), está direccionada para o intercâmbio de informação desportiva: resultados desportivos, horários e estatísticas de um grande número de competições. A norma utiliza XML para definir o conteúdo e estruturação de dados desportivos, facilitando a integração de serviços de informação desportiva de uma forma normalizada e rápida. O objectivo é a disponibilização mais rápida de resultados desportivos do que através de normas e formatos privados.

VAML

A norma VAML (*Video Annotation Markup Language*) surgiu no âmbito de aplicações de Hipervídeo (*Hypervideo*) (Zhou & Jin, 2004). Baseada na norma SGML, o VAML define um vídeo como um conjunto de elementos, marcas, atributos e entidades. Os elementos referem conteúdos do vídeo como sejam cenas, imagens, objectos e *hiperlinks*. As marcas contêm informação adicional inserida nos conteúdos vídeo com o objectivo de lhes adicionar um contexto.

TV-Anytime

A norma *TV-Anytime* (TV-Anytime, 2003) é uma especificação dedicada à entrega de conteúdos multimédia em gravadores de vídeo digital. O objectivo é o de permitir aos utilizadores experiências de televisão personalizadas. Esta normalização permite que os utilizadores acedam a conteúdos de uma grande variedade de fontes, adaptando-os às suas necessidades e preferências pessoais.

3.3 Mudança de Paradigma de Anotação

Nem as abordagens totalmente automáticas nem as predominantemente manuais satisfazem os requisitos actuais para um sistema de anotação genérico e eficiente.

Na abordagem de anotação automática há meios de extracção de características de conteúdo que obtêm descritores ditos de baixo nível semântico. A abordagem manual, por outro lado, fornece descritores de nível semântico alto. Os utilizadores, habitualmente, pretendem efectuar interrogações segundo conceitos de nível semântico alto. Da eventual desadequação entre o nível dos descritores e o nível utilizado nas interrogações surge o já referido fosso semântico (Sebe *et al.*, 2003). Um repositório rico em anotação manual permite um pesquisa mais próxima dos conceitos do domínio dos utilizadores.

Em relação aos métodos automáticos, os métodos manuais têm como desvantagens a morosidade do processo de anotação e a sua dependência em relação ao anotador. O aumento da eficácia do processo de anotação manual também passa pela redução significativa dos seus custos, objectivo que só se consegue alcançar com estratégias de reutilização de anotações, descrições e bases de conhecimento construídas previamente. A redução de custos da anotação manual pode ser conseguida tanto através da reutilização de metainformação existente em arquivos multimédia, como através de sistemas colaborativos de anotação já referidos que tiram partido dos novos fenómenos da *Web* como as Redes Sociais Online (Sack & Waitelonis, 2006) de que o YouTube (Hurley *et al.*, 2006; Wikipedia, 2007w), o Flickr (flickr, 2006; Wikipedia, 2006b) e del.icio.us (del.icio.us, 2006; Wikipedia, 2006a) são exemplos.

Propõe-se neste trabalho a utilização de técnicas de recuperação de imagem e vídeo como forma de melhoria de desempenho do processo de anotação manual. O objectivo é a utilização das ferramentas já existentes nas diversas áreas de recuperação de informação visual baseada em conteúdo para localização de segmentos vídeo que partilhem semelhanças em descritores de baixo nível. As anotações existentes para um segmento de vídeo podem ser propostas de forma semi-automática como descrições de outros segmentos semelhantes. A anotação manual serve-se de um processo de recuperação que visa encontrar informação com alto nível semântico para descrever o conteúdo multimédia em questão. A informação não apropriada pode, também deste modo, ser facilmente descartada durante o processo de anotação.

3.4 Proposta de Sistema de Reutilização de Anotações

O sistema proposto é um misto entre uma interface de anotação e um sistema de recuperação de informação baseada no conteúdo, integrando as funcionalidades de anotação e pesquisa de conteúdos numa mesma interface.

O sistema é composto por uma interface gráfica que permite ao utilizador inserir anotações sobre conteúdos à medida que efectua interrogações baseadas nestes, e por um repositório central que armazena tanto os conteúdos como a respectiva metainformação de alto valor semântico. Sobre o repositório operam vários serviços:

- um serviço de indexação vídeo que efectua segmentação;
- um serviço de extracção de características de baixo nível;
- um serviço de indexação que gera os índices a partir da metainformação obtida;
- um serviço de pesquisa e algoritmos de ordenação de resultados que possam ser relevantes para o excerto em questão. As interrogações pode ser efectuada sobre excertos pertencentes ao mesmo segmento vídeo ou a segmentos externos.

Deste processo de anotação/pesquisa e pesquisa/anotação resultam não só a reutilização de conhecimento mas também a construção de novas bases de conhecimento. Através de passos iterativos e incrementais, um utilizador fornece uma realimentação de alto valor semântico.

A combinação de técnicas de recuperação textual com técnicas baseadas em conteúdo fornece um processo interactivo capaz de gerar relações baixo nível/alto nível que tendem a minimizar o fosso semântico. Neste processo interactivo, as novas anotações introduzidas pelo utilizador e as relações com outro material multimédia completam e enriquecem o segmento actual através de uma realimentação de conhecimento para segmentos relacionados. A diminuição do fosso semântico é assim conseguida através das contribuições do anotador humano.

O sistema visa permitir não só a reutilização de metainformação previamente adicionada a outros conteúdos existentes no repositório, mas também utilizar o conhecimento do utilizador para agrupar características de baixo nível (descritores de cor, forma, textura, movimento) a informação de grande valor semântico (descrições de objectos,

pessoas, temas), independentemente de essa informação se encontrar ou não estruturada segundo uma norma específica. Por outras palavras, a eficácia do processo de reutilização de anotações é independente do modelo de estruturação de dados escolhido.

A reutilização de anotações pode ser feita utilizando semelhanças entre características de baixo nível e através de palavras-chave da anotação. A pesquisa e recuperação de material audiovisual relevante para reutilização de anotações pode ser feito através de técnicas de recuperação baseadas em texto, baseadas em conteúdo ou técnicas mistas.

3.5 Cenários de Utilização

Foi já referido que a reutilização de anotações pode ser efectuada no âmbito do mesmo segmento vídeo ou em segmentos de vídeo externos. Os casos onde a reutilização de anotações evidencia melhorias claras em relação ao um sistema tradicional de anotação manual são:

- **Variações de Conteúdo** - Devido a requisitos de visualização, transmissão e arquivo, é comum haver variações do mesmo conteúdo vídeo: mudanças em taxas de compressão, tipos de compressão, resolução espacial, espaço de cor (escala de cinzas), ou mesmo versões não sonorizadas. Esses diferentes formatos partilham a mesma informação semântica e consequentemente podem partilhar anotações ou mesmo metainformação independentemente das variações. A anotação de um segmento pode ser utilizada directamente em segmentos com características de baixo nível iguais (mesmo segmento vídeo mas com resolução ou codificação diferente).

- **Repetição de Cenas** - Frequentemente um segmento vídeo contém cenas repetidas ou imagens, objectos ou pessoas repetidas ao longo da sua duração. A detecção dessas repetições de cenas pode levar à propagação de anotações feitas posteriormente para os segmentos de vídeo similares, sejam eles na mesma fonte vídeo ou em fontes externas. As anotações relativas às cenas onde aparece um interveniente podem também ser propagadas entre cenas. Do ponto de vista de características de baixo nível (por exemplo detecção de faces) podem encontrar-se diversas cenas com a mesma pessoa e então a informação de alto nível (por

exemplo o nome) pode ser replicado para todos os segmentos com características de baixo nível semelhantes.

- **Relacionamento entre Segmentos** - Numa colecção há muitos segmentos vídeo que são implicitamente acerca de conteúdos semelhantes. A utilização de um processo interactivo de recuperação e anotação induz a geração de conhecimento através de técnicas de avaliação de similaridade por parte de um utilizador. A interacção humana é preciosa no estabelecimento de relações entre características de baixo nível e alto nível, conteúdos e anotações.

Capítulo 4

Sistema de Anotação Baseada em Pesquisa

O conceito de reutilização de anotações está intrinsecamente ligado a um processo de anotação baseada numa pesquisa. Ou seja, para haver uma reutilização de informação é necessário haver uma pesquisa prévia que evidencie segmentos de vídeo onde possa ser retirada informação. O processo de reutilização de anotações serve-se de um sistema de anotação que mapeia informação introduzida manualmente num determinado modelo de estruturação de informação, e também de um sistema de recuperação de informação, usado para pesquisar e recuperar informação relevante.

O sistema de anotação baseado em pesquisa é composto por uma interface de utilizador que serve para inserir anotações e efectuar pesquisas baseadas em conteúdos ou segundo palavras-chave, e de um serviço de recuperação de informação que opera sobre um repositório de conteúdos multimédia. A este serviço estão associadas as tarefas relativas a um comum sistema de recuperação baseada em conteúdos como são a segmentação de imagem, obtenção de descritores, construção de índices, processamento de interrogações e ordenação das respostas. O armazenamento das anotações, inseridas pelos utilizadores no módulo de anotação, é também uma responsabilidade deste serviço.

No capítulo anterior foi apresentado o conceito de anotação baseada em reutilização de informação. Neste capítulo, Secção 4.1, são abordados aspectos mais técnicos do processo de recuperação de informação visual, na Secção 4.2 é apresentada uma possível arquitectura do sistema e um ambiente experimental para prova de conceito na

Secção 4.3. Por último, as Secções 4.4 e 4.5 apresentam os resultados obtidos pelo ambiente experimental, e resumem algumas conclusões e considerações acerca da validade e desempenho do ambiente experimental. Esta análise tem como objectivos verificar a viabilidade da reutilização de descritores de baixo nível para identificação de similaridades e extracção de conhecimento.

4.1 Serviço de Recuperação de Informação

O serviço de recuperação de informação é o núcleo do sistema, é responsável pela gestão dos conteúdos e metainformação relacionada, e pela resposta aos pedidos do processo de anotação. Este módulo baseia-se no MPEG-7 XM e utiliza XML segundo a norma MPEG-7 para guardar metainformação de conteúdos e anotações. Este serviço é ainda responsável pela segmentação de imagem, extracção de descritores de baixo nível semântico e construção, com base nestes, dos índices que suportam a recuperação.

4.1.1 Estruturação de Informação

A escolha da norma de estruturação de informação para um sistema de recuperação de informação é um passo importante. No caso do sistema pretendido havia vários requisitos que deveriam ser satisfeitos sendo um dos principais a estruturação de metainformação de conteúdo (características de cor, forma, textura, movimento) de forma a facilitar a tarefa de pesquisa. Outras facilidades, como a extensibilidade a outros modelos de dados e a existência de implementações de extractores de características com mapeamento de descritores de conteúdos, foram factores adicionais que condicionaram a escolha tecnológica.

O modelo para estruturação de informação escolhido foi o MPEG-7. De entre as várias vantagens oferecidas pela norma, as que pesaram decisivamente para a escolha foram:

- **Tecnologia Baseada em XML** - A extensibilidade e compatibilidade do sistema com repositórios existentes era um dos requisitos. A norma MPEG-7 ao ser totalmente baseada em XML e definida através de um esquema XML (XSD - *XML*

Schema Definition) (Walmsley & Fallside, 2004; Beech *et al.*, 2004; Biron & Malhotra, 2004) satisfaz este requisito.

- **Suporte de Características de Conteúdo** - Era indispensável a utilização de uma norma que além de suportar descritores de conteúdos, por exemplo através de uma extensão ao seu modelo de dados, tivesse esses descritores de conteúdos normalizados. O MPEG-7 oferece através dos Esquemas de Descrição (*Description Schemes*) e Descritores (*Descriptors*) (Salembier & Smith, 2001) formatos normalizados para um grande número de características de baixo nível.

- **Algoritmos de Extracção de Características** - A extracção de características é uma tarefa pesada para a qual existem muitos algoritmos propostos. Estando fora de âmbito a realização de trabalho nesse campo, a existência de algoritmos para extracção de características disponíveis livremente foi também uma condicionante. O comité ISO do MPEG-7, através dos *Core Experiments* (Ohm *et al.*, 1999), promoveu a implementação no MPEG-7 XM de algoritmos para vários descritores.

- **Avaliação de Resultados** - A qualidade dos algoritmos de extracção de características é um factor decisivo na prestação de um sistema de recuperação. Nesse sentido a escolha do MPEG-7 tem também argumentos a favor. A comparação de várias implementações alternativas é feita por grupos especializados que tratam da avaliação de resultados obtidos por descritores e sistemas de recuperação baseados em MPEG-7. O exemplo de um desses grupos de avaliação é o TRECVID (NIST, 2003; Schulzrinne *et al.*, 2004).

Além das vantagens apresentadas, e no que respeita a metainformação descritiva dos conteúdos, o modelo de dados do MPEG-7 prevê anotações descritivas através dos descritores *FreeTextAnnotation*, *StructuredAnnotation*, *KeywordAnnotation*. A Tabela 4.1 ilustra um excerto XML MPEG-7 com essa informação. A descrições estrutural de conteúdos e informação acerca da criação e produção também pode ser mapeada no MPEG-7 através de descritores existentes, como *Title*, *Abstract* ou *Creator*. O suporte dos vários níveis de metainformação, de conteúdo, descritiva e estrutural, é um dos pontos mais fortes do MPEG-7.

O modelo de dados de suporte ao ambiente experimental demonstrador de conceito dá mais ênfase à metainformação de conteúdo (características de baixo nível) para recuperação de informação baseada em conteúdo. Ao nível de informação de descrição podem apenas ser utilizadas anotações de texto não estruturadas: texto livre

```
<Mpeg7>
    <Description xsi:type="ContentEntityType">
        <MultimediaContent xsi:type="VideoType">
            <Video>
                <TemporalDecomposition>
                    <MediaInformation id="FootBall">
                        <!-- ... -->
                    </MediaInformation>
                    <VideoSegment>
                        <TextAnnotation type="scene" relevance="1" confidence="1">
                            <\textbf{FreeTextAnnotation} xml:lang="en">
                            Zinedine Zidane scoring againstEngland.
                            </FreeTextAnnotation>
                            <KeywordAnnotation xml:lang="en">
                                <Keyword>Zinedine</Keyword>
                                <Keyword>Zidan</Keyword>
                                <Keyword>scoring</Keyword>
                                <Keyword>England</Keyword>
                            </KeywordAnnotation>
                            <StructuredAnnotation>
                                <Who>
                                    <Name xml:lang="en"> Zinedine Zidane </Name>
                                </Who>
                                <WhatAction>
                                    <Name xml:lang="en">
                                    Zinedine Zidane scoring against England.
                                    </Name>
                                </WhatAction>
                            </StructuredAnnotation>
                        </TextAnnotation>
                        <MediaTime>
                            <MediaTimePoint>T00:00:00:0F24</MediaTimePoint>
                            <MediaIncrDuration mediaTimeUnit="PT1N24F">1544</MediaIncrDuration>
                        </MediaTime>
                        <TemporalDecomposition>
                             <!-- ... -->
                        </TemporalDecomposition>
                    </VideoSegment>
                </TemporalDecomposition>
            </Video>
        </MultimediaContent>
    </Description>
</Mpeg7>
```

Tabela 4.1: Tipos de anotações descritivas MPEG-7

ou palavras-chave. Podem para o efeito ser utilizados os descritores *FreeTextAnnotation*, *StructuredAnnotation* ou *KeywordAnnotation*. A metainformação estrutural não é utilizada. Se por motivo de inclusão de novos requisitos de utilização podem ser adicionados outros esquemas de dados mais completos. O MPEG-7, através de extensões ao seu modelo de dados, é evolutivo e permite a inclusão de qualquer norma baseada em XML.

4.1.2 Segmentação de Imagem

O MPEG-7 XM é baseado em extracção de características para imagens estáticas. Para processamento de vídeo é necessária uma transformação de vídeo (independentemente do formato) para um tipo de imagens compatível com o MPEG-7 XM. O processo de segmentação de vídeo passa por dois passos distintos: a extracção de imagens estáticas (*frames*) partindo de sequências de vídeo e a detecção de cortes de cenas (*scene cuts*).

Extracção de Imagens Estáticas

O único formato de vídeo suportado pelo ambiente experimental é MPEG-2, embora outros formatos de vídeo pudessem ser adicionados com a integração dos respectivos descodificadores. Para efectuar a extracção das imagens estáticas, a partir de vídeo MPEG-2, foi utilizado o descodificador MPEG-2 (*mpeg2decode*) do MSSG (*MPEG Software Simulation Group*) (MSSG, 2006). O *mpeg2decode* é utilizado para extrair uma sequência de imagens estáticas (*frames*) do vídeo para o formato de imagem PPM (Netpbm, 2003). As sequências vídeo utilizadas têm uma frequência de imagens (*frame rate*) 25hz, o que implica que a cada segundo de vídeo correspondem 25 imagens. Os extractores de características implementados pelo MPEG-7 XM analisam imagens no formato PPM segundo a norma P6 (binário), codificação do tipo 2 (RGB) e com tamanho de 720 por 576 pixels. O descodificador MPEG-2 deve fornecer as imagens estáticas neste formato. A opção *-o3* mostrada no comando seguinte identifica a extracção de imagens no formato PPM binário.

```
mpeg2decode -b bitstream.mpg -f -r -o3 FootBall_%d
```

A Tabela 4.2 mostra um exemplo de uma imagem no formato PPM segundo a norma P3 (texto).

```
FootBall_0001.ppm
-------------------------------------------------------------
P3 720 576 255
  0  0  0    0  0  0    0  0  0   15  0 15
  0  0  0    0 15  7    0  0  0    0  0  0
  0  0  0    0  0  0    0 15  7    0  0  0
 15  0 15    0  0  0    0  0  0    0  0  0
 -------------------------------------------------------------
```

Tabela 4.2: Exemplo de imagem no formato PPM

Como ferramenta de auxílio ao processamento de imagem é utilizado o *ImageMagick* (Still, 2005). Este módulo serve para efectuar o carregamento das imagens, fazer conversões entre modos de representação de espaço de cor (RGB para YUV, RGB para HSL ou outras) e fornecer uma interface de programação para que os extractores de características possam ter acesso directo aos pixels da imagem.

Detecção de Cortes de Cena

A detecção de cortes de cena é feita através do extractor de características *Video Editing* e do descritor *Video Editing* do MPEG-7 XM. Este extractor utiliza a detecção de pontos KLT (Lucas & Kanade, 1981) nas imagens e através de uma filtragem de Kalman (Kalman, 1960) e faz a previsão da posição dos pontos de características (*feature points*) para a imagem seguinte. Através de um valor de limiar são determinadas as mudanças significativas em pontos de características para que uma imagem seja marcada como sendo um corte de cena.

No XML resultante da extracção de características do *Video Editing*, os segmentos separados por dois cortes de cena são assinalados por nós XML `EditedVideoSegment`, enquanto que as imagens cortes de cena são assinalados por nós `Transition`. O excerto XML da Tabela 4.3 dá um exemplo de ambos os tipos de nós.

De modo a manter a relação entre os descritores extraídos e as imagens relacionadas é construída, durante o processo de segmentação, uma listagem de ficheiros de imagens. Esta listagem é utilizada para efectuar o mapeamento das sequências das imagens e os índices dos descritores. O exemplo na Tabela 4.4 ilustra a forma de um desses ficheiros.

Do exemplo do descritor mostrado na Tabela 4.3, e da listagem de imagens estáticas da Tabela 4.4, pode verificar-se que existe um segmento de vídeo `EditedVideoSegment` delimitado pelos índices `MediaRelIncrTimePoint` com valor 0 (imagem `OtherSideOfHeaven2_0.ppm`) e `MediaIncrDuration` valor 77 (0 + 77) (imagem `OtherSideOfHeaven2_77.ppm`). Também é mostrado um corte de cena `Transition` no índice `MediaRelIncrTimePoint` com valor 77 (imagem `OtherSideOfHeaven2_77.ppm`).

```
<Mpeg7>
    <DescriptionUnit xsi:type="DescriptorCollectionType">
        <VideoEditing use="analytic">
            <SegmentDecomposition gap="false"
                                  type="temporal"
                                  overlap="true"
                                  temporalConnectivity="true">
                <EditedVideoSegment id="EVS:0"
                                    use="analytic"
                                    editingLevel="shot">
                    <MediaTime>
                        <MediaRelIncrTimePoint timeUnit="PT1N30">
                        0
                        </MediaRelIncrTimePoint>
                        <MediaIncrDuration timeUnit="PT1N30">
                        77
                        </MediaIncrDuration>
                    </MediaTime>
                </EditedVideoSegment>
                <Transition id="Trans:0"
                            use="analytic"
                            evolution="cut"
                            editingLevel="global">
                    <MediaTime>
                        <MediaRelIncrTimePoint timeUnit="PT1N30">
                        77
                        </MediaRelIncrTimePoint>
                        <MediaIncrDuration timeUnit="PT1N30">
                        1
                        </MediaIncrDuration>
                    </MediaTime>
                </Transition>
            </SegmentDecomposition>
        </VideoEditing>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.3: Excerto do descritor *Video Editing* com indicação de nós `EditedVideoSegment` e `Transition`

4.1.3 Obtenção de Descritores

A obtenção de descritores é feita com recurso ao *software* MPEG-7 XM, configurado de forma a efectuar a extracção dos 6 descritores (*Scalable Color*, *Color Layout*, *Edge Histogram*, *Homogeneous Texture*, *Contour Shape* e *Dominant Color*) em paralelo. Esse processo é repetido para cada uma das imagens extraídas no processo de segmentação descrito em 4.1.2.

Através da conjugação dos módulos de Descritores (*Descriptors*) e Esquemas de Descrição (*Description Schemes*), Ferramentas de Extracção (*Extraction Tools*) e Esquemas de Codificação (*Coding Schemes*), forma-se a aplicação de extracção de características para o descritor em questão (ver Figura 2.11). Cada um desses módulos tem a sua função específica:

```
OtherSideOfHeaven2.lst
------------------------------------------------------------

D:\MyStreams\PPM-00000000-00006000\OtherSideOfHeaven2_0.ppm
D:\MyStreams\PPM-00000000-00006000\OtherSideOfHeaven2_1.ppm
D:\MyStreams\PPM-00000000-00006000\OtherSideOfHeaven2_2.ppm
D:\MyStreams\PPM-00000000-00006000\OtherSideOfHeaven2_3.ppm
D:\MyStreams\PPM-00000000-00006000\OtherSideOfHeaven2_4.ppm
(...)
D:\MyStreams\PPM-00000000-00006000\OtherSideOfHeaven2_77.ppm
(...)
------------------------------------------------------------
```

Tabela 4.4: Mapeamento de índices e listagem de ficheiros de imagens

- **Descritores** (*Descriptors*) **e Esquemas de Descrição** (*Description Schemes*) - Este módulo implementa a estrutura de dados normalizada dos Descritores e Esquemas de Descrição. Essas classes também fornecem métodos para aceder aos elementos normativos dos descritores. No caso do Esquema de Descrição Genérico (*GenericDS*) é fornecida, não uma estrutura de dados específica, mas uma interface à biblioteca XML para que possam ser utilizadas as árvores e estruturas definidas para o Descritor e Esquema de Descrição que é instanciado.

- **Ferramentas de Extracção** (*Extraction Tools*) - As Ferramentas de Extracção são específicas a cada um dos Descritores e Esquemas de Descrição. A implementação destas ferramentas não está normalizada, mas devem fornecer uma descrição válida. Estas ferramentas são responsáveis por analisar os conteúdos e fornecer os valores necessários para a construção dos Descritores.

- **Esquemas de Codificação** (*Coding Schemes*) - Neste módulo são definidos os esquemas de codificação e descodificação para cada um dos Descritores e Esquemas de Descrição. Os Esquemas de Codificação servem também para efectuar codificações de descritores para o formato binário (BiM - *MPEG-7 Binary Format for XML Data*) (Niedermeier *et al.*, 2002). Na presente abordagem todas as representações de Descritores são feitas em XML, por motivos de performance pode a qualquer altura optar-se pela codificação de descritores binária.

De seguida exemplificam-se excertos de XML representativos dos descritores utilizados (Cieplinski *et al.*, 2001):

Scalable Color

Este descritor (Tabela 4.5) contém a indicação do número de coeficientes utilizado na sua representação através do atributo `NumberOfCoefficients` do nó `Descriptor`, o número de *bitplanes* (Wikipedia, 2007a) descartados na representação de cada um dos coeficientes, e a lista dos coeficientes do descritor no nó `Coefficients`.

```
<Mpeg7>
    <DescriptionUnit xsi:type = "DescriptorCollectionType">
        <Descriptor xsi:type="ScalableColorType"
                    NumberOfCoefficients="64"
                    NumberOfBitplanesDiscarded="3">
            <Coefficients>
            -4  3  10  0  0  0  0  1  -3  1  3  0  0  1  -4  1  0 0  0  1  -1  0
            0  0  -1  0  -1  0  0  0  0  0  1  1  1  0  1  0  0  0  0  1  0  0 0
            0  0  0  0  0  0  0  -1  0  0  0  0  0  0  0  0  0  0  0
            </Coefficients>
        </Descriptor>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.5: Exemplo XML de descritor *Scalable Color*

Color Layout

Este descritor (Tabela 4.6) utiliza uma codificação de cor YCbCr (Wikipedia, 2007v) onde Y representa a componente da luminância e Cb e Cr as componentes de crominância para os canais azul e vermelho.

No descritor estão representados cada um dos primeiros coeficientes da DCT (*Discrete Cosine Transform*) (Ahmed *et al.*, 1974) quantificada para as componente luminância (Y), crominância azul (Cb) e crominância vermelha (Cr) utilizando os nós `YDCCoeff`, `CbDCCoeff` e `CbDCCoeff` respectivamente. No exemplo, os valores da DCT quantificada são representados nos nós `YACCoeff5`, `CbACCoeff2` e `CrACCoeff2`.

Edge Histogram

Neste descritor (Tabela 4.7) cada imagem é subdividida em 16 sub-imagens sendo para cada uma destas contabilizado o número de contornos segundo as 5 direcções (vertical, horizontal, 45º, 135º e não direccionais). No total o descritor é composto por 80 (16x5) coeficientes agrupados no nó `BinCounts`.

```
<Mpeg7>
    <DescriptionUnit xsi:type = "DescriptorCollectionType">
        <Descriptor xsi:type="ColorLayoutType">
            <YDCCoeff>22</YDCCoeff>
            <CbDCCoeff>40</CbDCCoeff>
            <CrDCCoeff>25</CrDCCoeff>
            <YACCoeff5>21 15  20  12  15 </YACCoeff5>
            <CbACCoeff2>19  21 </CbACCoeff2>
            <CrACCoeff2>13  13 </CrACCoeff2>
        </Descriptor>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.6: Exemplo XML de descritor *Color Layout*

```
<Mpeg7>
    <DescriptionUnit xsi:type = "DescriptorCollectionType">
        <Descriptor xsi:type="EdgeHistogramType">
            <BinCounts>
            0  1  0  0  0 0 3  0  0  0  0  5  0  0  0  1  3  1  1 0 0  5  1  1 0
            6  0 3 3 1  5  4  2  2  4  4  3  4  2  0  5  3 2 3  0  6  2  2 2 4 3
            2  2 3  5  2  3  2  3  1  5  4  5  5  1  5 4 4  4  3  4  5  3 5 6  3
            5 3  3
            </BinCounts>
        </Descriptor>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.7: Exemplo XML de descritor *Edge Histogram*

Homogeneous Texture

Este descritor (Tabela 4.8) inclui a média e desvio padrão da intensidade de todos os pixels da imagem nos nós `Average` e `StandardDeviation`. As componentes de energia e desvios padrão de cada uma das componentes de texturas encontram-se numa lista de coeficientes, nos nós `Energy` e `EnergyDeviation`.

Contour Shape

Neste descritor (Tabela 4.9) encontram-se discriminados os valores de características que captam a natureza dos contornos dos objectos representados na imagem. Os valores do nó `GlobalCurvature` representam as características de circularidade (*roundness*) e excentricidade (*eccentricity*). Depois de aplicada uma filtragem aos valores de circularidade e excentricidade, os respectivos coeficientes são representados no nó `PrototypeCurvature`. O descritor contém adicionalmente o valor máximo dos picos da filtragem no nó `HighestPeakY`, e restantes valores de picos, atributos `peakX` e `peakY` dos nós `Peak`.

```
<Mpeg7>
    <DescriptionUnit xsi:type = "DescriptorCollectionType">
        <Descriptor xsi:type="HomogeneousTextureType">
            <Average>113</Average>
            <StandardDeviation>6</StandardDeviation>
            <Energy>
            131  172  174  136 172  173  123  145  145  115  143  143  97  112
            97  103  92  100 68 60  58  74  56  45  78  12  9  40  7  3
            </Energy>
            <EnergyDeviation>
            127  175  176  134  175  175  112  145  145  110 138  141  85  109
            87  91  83  98  62  50  55  64  54  44  91  4  6 35  3  0
            </EnergyDeviation>
        </Descriptor>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.8: Exemplo XML de descritor *Homogeneous Texture*

```
<Mpeg7>
    <DescriptionUnit xsi:type = "DescriptorCollectionType">
        <Descriptor xsi:type="ContourShapeType">
            <GlobalCurvature>2  1 </GlobalCurvature>
            <PrototypeCurvature>1  1 </PrototypeCurvature>
            <HighestPeakY>9</HighestPeakY>
            <Peak peakX="60" peakY="4"/>
            <Peak peakX="0"  peakY="3"/>
            <Peak peakX="61" peakY="5"/>
            <Peak peakX="56" peakY="6"/>
            <Peak peakX="2"  peakY="4"/>
            <Peak peakX="1"  peakY="7"/>
            <Peak peakX="55" peakY="6"/>
            <Peak peakX="62" peakY="6"/>
            <Peak peakX="55" peakY="7"/>
        </Descriptor>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.9: Exemplo XML de descritor *Contour Shape*

Dominant Color

Este descritor (Tabela 4.10) contém informação acerca da coerência espacial da cor dominante no nó `SpatialCoherency`. A discretização das cores dominantes é feita através de uma lista de índices de cores e valores de percentagem de cobertura na área da imagem descrita na lista de nós `Value` e pares de nós filhos `Percentage` e `Index`.

4.1.4 Construção de Índices

A construção de índices é realizada com o agrupamento num único documento XML dos resultados da extracção dos descritores *Video Editing*, *Scalable Color*, *Color Layout*,

```
<Mpeg7>
    <DescriptionUnit xsi:type = "DescriptorCollectionType">
        <Descriptor xsi:type="DominantColorType">
            <SpatialCoherency>0</SpatialCoherency>
            <Value>
                <Percentage>20</Percentage>
                <Index>3847  0  0 </Index>
            </Value>
            <Value>
                <Percentage>2</Percentage>
                <Index>0  2691213 -8126464 </Index>
            </Value>
            <Value>
                <Percentage>6</Percentage>
                <Index>1564672  0  -291034 </Index>
            </Value>
            <Value>
                <Percentage>1</Percentage>
                <Index>0  0  1 </Index>
            </Value>
            <Value>
                <Percentage>1</Percentage>
                <Index>0  0  0 </Index>
            </Value>
        </Descriptor>
    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.10: Exemplo XML de descritor *Dominant Color*

Edge Histogram, *Homogeneous Texture*, *Contour Shape* e *Dominant Color*. Por forma relacionar os descritores com as imagens respectivas é necessário utilizar a tabela que contém os nomes das imagens ordenadas (Tabela 4.4). Dessa forma consegue saber-se que o 3º grupo de descritores corresponde à 3ª imagem do repositório.

O XML exemplo da Tabela 4.11 mostra a estruturação por índices dos documentos XML contendo os descritores.

A correlação entre os descritores e as imagens é feita através do índice do descritor e número de linha do respectivo ficheiro de listagem.

4.1.5 Pesquisa em Índices e Metodologias de Ordenação

Para efectuar pesquisas é necessário utilizar os índices construídos e processar, com as ferramentas utilizadas para o repositório, a imagem ou imagens a pesquisar (interrogação). O processamento da interrogação faz-se com as ferramentas de extracção já descritas. Na pesquisa sobre os índices são utilizado dois módulos do MPEG-7 XM, um para calcular as distâncias entre os descritores e outro que efectua uma ordenação das imagens com maior similaridade do ponto de vista das características (ver Figura

```
<Mpeg7>
    <DescriptionUnit xsi:type="DescriptorCollectionType">
        <VideoEditing use="analytic">
            <!-- Conteúdo do Descritor de Detecção de Cortes de Cena -->
        </VideoEditing>

        <!-- Bloco de Descritores para a 1ª Imagem -->
        <Descriptor xsi:type="ContourShapeType">
            <!-- ... -->
        </Descriptor>
        <Descriptor xsi:type="ScalableColorType"
            <!-- ... -->
        </Descriptor>
        <Descriptor xsi:type="ColorLayoutType">
            <!-- ... -->
        </Descriptor>
        <Descriptor xsi:type="EdgeHistogramType">
            <!-- ... -->
        </Descriptor>
        <Descriptor xsi:type="HomogeneousTextureType">
            <!-- ... -->
        </Descriptor>
        <Descriptor xsi:type="DominantColorType">
            <!-- ... -->
        </Descriptor>

        <!-- Bloco de Descritores para a 2ª Imagem -->
        <Descriptor xsi:type="ContourShapeType">
            <!-- ... -->
        </Descriptor>

        <!-- ... -->

        <!-- Bloco de Descritores para a n-esima Imagem -->

    </DescriptionUnit>
</Mpeg7>
```

Tabela 4.11: Representação de excerto XML de descritores para segmento vídeo

2.12):

- **Ferramentas de Pesquisa** (*Search Tools*) - Cada tipo de descritor necessita de um algoritmo diferente para cálculo de distâncias de similaridade. Este módulo inclui a implementação dos algoritmos de medidas de distância respectivo a cada um dos descritores. Através de duas instâncias de descritores (o descritor da imagem de interrogação e um descritor de imagem da base de dados), o módulo retorna um valor de distância entre as duas imagens baseada no descritor. O processo de pesquisa percorre todos os índices construindo uma lista de valores de distâncias.

- **Ordenação** (*Results*) - Neste módulo é feita a integração de todos os resultados de distâncias dos vários descritores. São aplicadas correcções ao valores obtidos

(factor correctivo associado à importância do descritor), e é efectuada uma ordenação decrescente por valor de semelhança. Os valores de similaridade variam entre 0 para uma similaridade total e 255 no caso de haver uma dissimilaridade total.

Os algoritmos de similaridades disponíveis no módulo de Ferramentas de Pesquisa (*Search Tools*) do MPEG-7 XM para cálculo do valores de distância entre imagens são os seguintes (Jeannin, 1999):

Scalable Color

Existem 3 modos distintos para cálculo de similaridade do descritor *Scalable Color*: comparação de histogramas através da reconstrução do histograma de cor, cálculo de semelhança no domínio dos coeficientes de Haar (Haar, 1910) e distância de Hamming (Exoo, 2003).

A reconstrução do histograma de cor partindo dos coeficientes da Haar permite um ganho na eficácia e precisão no cálculo do valor de semelhança comparativamente à utilização directa dos coeficientes. A comparação de histogramas é um método viável apenas quando se pretende resultados com elevada precisão, ou seja deve ser utilizada apenas quando todos os coeficientes estão disponíveis.

O cálculo do valor de semelhança no domínio dos coeficientes de Haar é feito com recurso à normalização L1 (Horn & Johnson, 1990). A utilização directa dos coeficientes de Haar introduz um pequeno erro na precisão do valor de semelhança mas permite diminuir significativamente o custo computacional.

No caso de apenas serem guardados os sinais dos coeficientes, o cálculo é feito através da distância de Hamming. Esta distância é calculada através da comparação directa de dois descritores de 63 bits (o operador XOR, e soma de bits 1 do resultado, é uma forma de implementar essa distância).

Color Layout

A similaridade entre os valores de dois descritores *Color Layout* $CLD_1(Y_1, Cb_1, Cr_1)$ e $CLD_2(Y_2, Cb_2, Cr_2)$ é calculada utilizando a seguinte formula de distância

$$
\begin{aligned}
dist(CLD_1, CLD_2) = & \sqrt{\sum_{i=0}^{n-1} \lambda_{Yi} \left(Y_1[i] - Y_2[i]\right)^2} \\
& + \sqrt{\sum_{i=0}^{n-1} \lambda_{Cbi} \left(Cb_1[i] - Cb_2[i]\right)^2} \\
& + \sqrt{\sum_{i=0}^{n-1} \lambda_{Cri} \left(Cr_1[i] - Cr_2[i]\right)^2}
\end{aligned} \tag{4.1}
$$

onde n é o máximo entre o número de coeficientes dos descritores das duas imagens, λ_{Yi}, λ_{Cbi} e λ_{Cri} são pesos a atribuir a cada coeficiente conforme especificado por Jeannin (1999) na parte normativa do MPEG-7 XM.

Edge Histogram

A medida de similaridade entre dois descritores *Edge Histogram* EDH_1 e EDH_2 é dada pela seguinte distância

$$
\begin{aligned}
dist(EDH_1, EDH_2) = & \sum_{i=0}^{79} |LocalEdge_1[i] - LocalEdge_2[i]| \\
& + 5 \times \sum_{i=0}^{4} |GlobalEdge_1[i] - GlobalEdge_2[i]| \\
& + \sum_{i=0}^{64} |SemiGlobalEdge_1[i] - SemiGlobalEdge_2[i]|
\end{aligned} \tag{4.2}
$$

onde $LocalEdge_1[i]$ e $LocalEdge_2[i]$ representam os valores $BinCount[i]$ das imagens 1 e 2 reconstruídos através da aplicação das tabelas de quantificação da parte normativa do descritor *Edge Histogram*. $GlobalEdge_1[i]$ e $GlobalEdge_2[i]$ representam os valores decimais para os histogramas de contorno, enquanto que $SemiGlobalEdge_1[i]$ e $SemiGlobalEdge_2[i]$, representam os valores do histograma $SemiGlobalEdge$. Uma vez que o número de valores dos histogramas globais $GlobalEdge_1[i]$ e $GlobalEdge_2[i]$ são relativamente inferiores aos locais ($LocalEdge_1[i]$ e $LocalEdge_2[i]$) e semi-globais

($SemiGlobalEdge_1[i]$ e $SemiGlobalEdge_2[i]$) é-lhe aplicado um peso de 5.

Homogeneous Texture

A distância de similaridade entre dois descritores de textura *Homogeneous Texture*, HTD_1 e HTD_2, é calculada pela soma das diferenças absolutas dos valores dos descritores aplicando um factor de normalização (α)

$$dist(HTD_1, HTD_2) = \sum_{i=0}^{k} \left| \frac{HTD_1(k) - HTD_2(k)}{\alpha(k)} \right| \tag{4.3}$$

O factor de normalização normalmente utilizado para $\alpha(k)$ é o valor de desvio padrão de $HTD_2(k)$.

Contour Shape

O cálculo de similaridade entre dois descritores *Contour Shape* CSD_1 e CSD_2 é bastante complexo computacionalmente. A medida de similaridade é processa-se em duas fases, a primeira calcula um valor aproximado da semelhança dos dois contornos. Caso eles sejam significativamente diferentes são marcados como não similares e não é apurado qualquer valor de semelhança. Para os restantes casos é utilizada a seguinte medida

$$M = 0.4 \times \frac{\mid CSD_1[0] - CSD_2[0] \mid}{max(CSD_1[0], CSD_2[0])} + 0.3 \times \frac{\mid CSD_1[1] - CSD_2[1] \mid}{max(CSD_1[1], CSD_2[1])} + M_{CSS} \tag{4.4}$$

A similaridade entre dois conjuntos de picos do CSS (*Curvature Scale Space*) (Mokhtarian & Bober, 2003) é essencialmente uma normalização L2 (Horn & Johnson, 1990) entre os picos coincidentes com uma penalização para os não coincidentes. Dois picos dizem-se coincidentes se a distância L2 entre as suas coordenadas x está abaixo de um determinado limiar (*threshold*).

A medida de similaridade é dada por

$$M_{CSS} = \sum_{1}((xpeak[i] - xpeak[j])^2 + (ypeak[i] - ypeak[j])^2) + \sum_{2}(ypeak[i])^2 \quad (4.5)$$

em que $\sum_1$ é a soma de todos os picos coincidentes e $\sum_2$ a soma dos não coincidentes.

Dominant Color

A distância de similaridade entre dois descritores *Dominant Color* DCD_1 e DCD_2 é calculado como

$$dist(DCD_1, DCD_2) = W1 \times \textit{SCDiff} \times \textit{DCDiff} + W2 \times \textit{DCDiff} \quad (4.6)$$

onde

SCDiff $= abs(SpatialCoherency_1 - SpatialCoherency_2)$, $SpatialCoherency_1$ e $SpatialCoherency_2$ são normalizados para um intervalo entre 0 e 1. E quantificados não uniformemente de forma a calcular *SCDiff*. *DCDiff* é a diferença entre dois conjuntos de cores dominantes, $W1$ e $W2$ são os pesos do 1º e 2º termo respectivamente.

4.2 Arquitectura

O Sistema de anotação baseada em pesquisa segue uma arquitectura cliente-servidor. O servidor (Figura 4.1) é um sistema de recuperação de vídeo típico juntamente com um sistema de recuperação de texto. Por sua vez, a aplicação cliente (Figura 4.2) é uma interface gráfica onde o utilizador introduz novas anotações e pesquisa material relevante para reutilização de anotações.

4.2.1 Servidor

A arquitectura do servidor inclui o sistema de recuperação de imagem/vídeo e o repositório vídeo, onde são armazenados os itens de vídeo e respectivos índices.

O servidor suporta dois tipos de processos:

- **Indexação** - Neste processo incluem-se a segmentação das cenas, a obtenção de descritores MPEG-7 e a construção de índices;
- **Pesquisa** - Este processo é responsável por efectuar interrogações sobre o repositório (pesquisa por exemplos ou textual) e recuperar os documentos relevantes.

Na Figura 4.1 podem observar-se 3 módulos: indexação (*Indexing*), interrogação (*Query*) e ordenação (*Results*). O servidor interage com a interface de utilizador através dos módulos interrogação e ordenação. No módulo de interrogação o servidor recebe informação dos descritores que compõem uma interrogação, no módulo de ordenação o servidor fornece à interface de utilizador listas de documentos ordenadas por relevância. O servidor tem ainda como entrada documentos vídeo que guarda no repositório e fornece ao módulo de indexação para serem processados, extraindo os respectivos descritores MPEG-7. Este módulo (módulo de indexação) contém cada um dos extractores de descritores que podem ser adicionados ou removidos ao sistema. Adicionalmente aos conteúdos de vídeo, o repositório também guarda os documentos XML contendo os descritores MPEG-7 dos conteúdos.

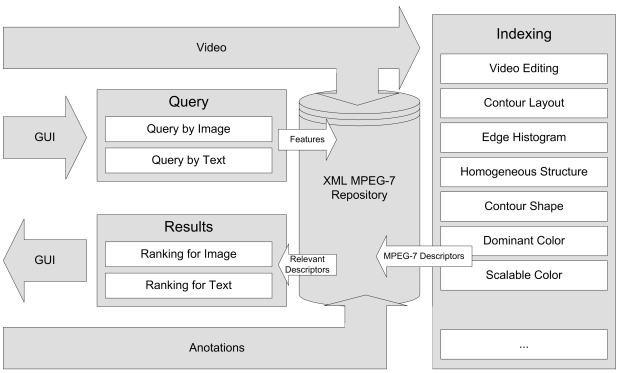


Figura 4.1: Arquitectura do servidor

4.2.2 Interface de Utilizador

A interface de utilizador interage com o servidor importando novos conteúdos audiovisuais para o sistema. Também fornece métodos de guia ao utilizador na anotação de novos conteúdos servindo-se de um processo recuperação (recuperação baseada em

conteúdo ou textual). Aqui são obtidas referências a documentos cuja metainformação pode ser reutilizada para anotação de novos conteúdos vídeo.

A Figura 4.2 mostra um painel de anotação onde se encontram um reprodutor vídeo (*player*) e uma interface de inserção de metainformação. Também inclui um painel para importação de vídeo no sistema. Existem também dois painéis que fazem a interface aos módulos de interrogação e ordenação do servidor. O painel de metainformação relevante (*Relevant Metadata Panel*) comanda ambos os módulos de interrogação e ordenação por forma a que o processo de anotação com reutilização de anotação seja transparente para o utilizador. Neste painel são apresentados possíveis excertos de vídeo cuja metainformação pode ser reutilizada.

A interface de utilizador suporta dois processos que envolvem interacção com o repositório MPEG-7:

- **Anotação** - Neste processo o utilizador adiciona aos conteúdos audiovisuais anotações ou metainformação de forma manual;
- **Recuperação** - Pesquisa de anotações ou metainformação relevante no repositório, que possa ser relevante para o conteúdo multimédia a ser anotado.

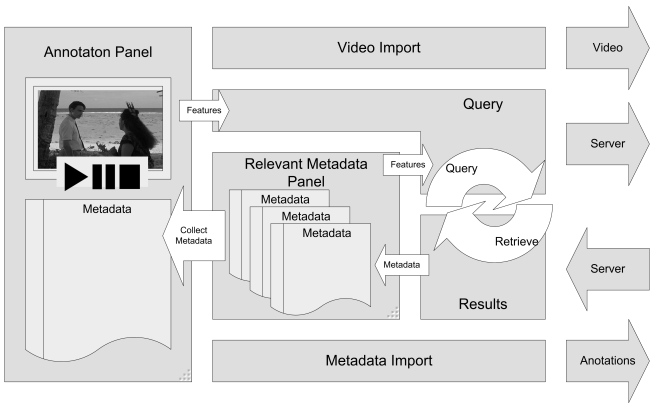


Figura 4.2: Arquitectura da interface de utilizador

4.3 Ambiente Experimental

Por forma a efectuar uma validação da proposta de anotação baseada em pesquisa foi elaborado um ambiente experimental de teste. Neste ambiente fez-se análise de con-

teúdo a vários segmentos vídeo com o intuito de encontrar similaridades entre diferentes cenas através da extracção e comparação de descritores. Sendo esse o ponto fulcral do sistema de anotação baseada em pesquisa, uma não adequação dos descritores e métricas de similaridade invalidaria a detecção de cenas semelhantes e consequentemente a reutilização de metainformação por parte do sistema de anotação assistida.

Os descritores MPEG-7 do XM são muito utilizados em sistemas experimentais de recuperação de imagem baseada em conteúdo (CBIR), o que os tornou apropriados para a elaboração deste ambiente experimental. Para avaliar o desempenho e adequação dos descritores seleccionados às tarefas de apoio à anotação propostas, foi necessário efectuar trabalho experimental em duas áreas: na detecção de cortes de cena e no cálculo de similaridades entre imagens. Ambas são importantes. A detecção de cortes de cenas é importante no processo de reutilização de anotações uma vez que detecta cortes de cena úteis na identificação dos diferentes segmentos. Por outro lado, o processo de similaridades entre imagens é central no cálculo de medidas de distâncias entre imagens pertencentes a segmentos de vídeo, ou seja, as distâncias entre descritores de imagens pertencentes a cenas diferentes que servem para identificar cenas similares.

4.3.1 Processo

Para efectuar trabalho experimental foram escolhidos os descritores *Scalable Color*, *Color Layout*, *Edge Histogram* e *Homogeneous Texture*, em conjunto com os respectivos extractores de características. Foram deixados de fora os descritores *Contour Shape* e *Dominant Color*. O descritor *Video Editing* e seu extractor foi utilizado no âmbito de detecção de cortes de cena.

Nesta secção, para ilustrar todo o processo, os exemplos dados recaem apenas num único segmento vídeo (excerto vídeo *Other Side Of Heaven*). São também utilizados apenas resultados obtidos pelo descritor *Scalable Color* e cortes de cena obtidos pelo descritor *Video Editing*. Na Secção 4.4 são apresentados resultados para os restantes descritores utilizando um conjunto de segmentos vídeo mais alargado.

O processo utilizado durante a obtenção de resultados experimentais divide-se em quatro partes:

- **Extracção de Imagens** - A extracção de imagens estáticas para posterior análise e cálculo de descritores é efectuada através do descodificador de vídeo MPEG-

2 da (MSSG, 2006). Do segmento vídeo em questão resultam um conjunto de imagens (1 por *frame* de vídeo) no formato PPM (Netpbm, 2003). Este processo já documentado na Secção 4.1.2 é o passo inicial da análise de segmentos de vídeo para detecção de similaridades;

- **Extracção de Características** - À semelhança do processo utilizado na Secção 4.1.2, as imagens obtidas são posteriormente analisadas por uma aplicação (`MPEG7ExtractDemo`) desenvolvida no âmbito do ambiente experimental. Esta aplicação está baseada no MPEG-7 XM e tem a particularidade de extrair diverso descritores em simultâneo. Da análise de cada uma das imagens pelos diversos extractores de características, *Scalable Color*, *Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Video Editing* do MPEG-7 XM, resulta um ficheiro XML (`MultiExtractorDescriptors.xml`) contendo os vários descritores para cada uma das imagens. A Tabela 4.11 ilustra o formato deste ficheiro XML;

- **Detecção de Cortes de Cena** - O cálculo das distâncias par a par entre os descritores das imagens fornece informações de cortes de cena e similaridades entre segmentos. O descritor *Video Editing* fornece a listagem dos cortes de cena para o segmento vídeo. A Secção 4.1.2 ilustra esse processo. Outro método que pode ser utilizado para detecção de cortes de cena é a análise de semelhanças entre descritores de imagens consecutivas. Caso os descritores tenham uma elevada similaridade pode considerar-se que ambas as imagens pertencem a uma mesma cena. Caso os descritores sejam dissimilares pode considerar-se que se trata de um corte de cena. Para efeitos de ambiente experimental apenas será feita a validação do algoritmo de cortes de cena do descritor *Video Editing*;

- **Cálculo de Similaridade entre Imagens Par a Par** - Duas imagens consecutivas, extraídas de um segmento de vídeo cujas características de baixo nível sejam semelhantes, têm uma elevada probabilidade de pertencem a uma mesma cena. Por outro lado, quando as características das duas imagens consecutivas são diferentes, supõe-se a presença de duas cenas diferentes. Partindo deste pressuposto foi desenvolvida uma aplicação (`MPEG7RankDemo`) que calcula as semelhanças par a par entre todas as imagens do segmentos vídeo para os descritores *Scalable Color*, *Color Layout*, *Edge Histogram* e *Homogeneous Texture*. Para cada descritor e par de imagens é calculado um valor de semelhança. Do cálculo de todas as semelhanças entre as imagens do segmento resultam mapas de semelhança como representado na Figura 4.3. Nestes mapas podem ser observadas zonas de

imagens com características semelhantes e por outro lado zonas de fronteira que representam zonas de corte de cena. De referir que os valores de semelhança fornecem valores absolutos próximos de zero para imagens similares, e valores absolutos elevados para imagens dissimilares. A Secção 4.3.2 dá mais pormenor sobre a matriz de similaridades par a par.

- **Cálculo de Valor de Semelhança entre Cenas** - Uma cena vídeo é constituída por uma ou várias imagens. No ponto anterior foram calculadas as semelhanças entre todas as imagens de um determinado excerto vídeo. Para detecção de cenas similares é necessário construir uma matriz de todas as semelhanças de cenas par a par. Esta matriz é em todo semelhante à de similaridade de imagens par a par, diferindo apenas no facto de esta conter um agrupamento das similaridades de imagens delimitadas por dois cortes de cena. Deste agrupamento das similaridades surge um valor de similaridade que é representativo de todas as similaridades de imagens contidas na cena (similaridades das imagens entre os dois cortes de cena). No cálculo do valor de similaridade representativo da cena (com vários valores de semelhança de imagens par a par) foram considerados 3 cálculos aritméticos simples: valor Máximo, Mínimo e Médio dos valores de similaridades das imagens que constituem a cena.

- **Avaliação de Semelhança entre Cenas** - Para efectuar a recuperação de cenas semelhantes devem ser tomadas as seguintes considerações: O Máximo representa o valor de semelhança entre imagens contidas na cena que obteve o valor de dissimilaridade mais elevado. Logo valores de Máximo baixos revelam elevadas probabilidades de as cenas serem similares. O Mínimo representa o valor de semelhança entre imagens contidas na cena que obteve o valor de similaridade mais elevado. Um valor de Mínimo elevado revela que as cenas mais semelhantes têm um valor de semelhança baixo, logo a probabilidade de pertencerem a cenas semelhantes também é baixo. O valor de Média uniformiza as discrepâncias entre valores de Máximo e Mínimo, se este valor for próximo do Máximo e do Mínimo é de esperar uma cena com alguma uniformidade. O facto destes valores serem muito afastados indica que é uma cena com alguma irregularidade de semelhança entre as imagens que a compõem, devendo ser efectuada uma avaliação de cenas Similares com mais precaução. O método de avaliação adoptado foi o de escolher para uma determinada cena as cenas que apresentam os melhores valores de semelhança para valores Máximo, Médio e Mínimo das semelhanças entre as imagens que as compõem. Por forma a minimizar o tamanho da resposta

relativa à detecção de cenas similares foram escolhidos 3 métodos: as 5 cenas que apresentam maiores valores de similaridade par a par, as 5 cenas que apresentam maiores valores de similaridade par a par acima de um limiar de semelhança pré-definido, e todas as cenas com valores de semelhança par a par acima do limiar de semelhança. O primeiro retorna sempre as 5 cenas com melhores valores de semelhança, independentemente de terem um valor aceitável de similaridade. O segundo retorna apenas as 5 cenas se elas tiverem um valor mínimo de similaridade definido pelo limiar de semelhança. Por último, a resposta é dada por todas as cenas com valores de semelhança acima do limiar.

Uma vez que os métodos utilizados durante o processo para Extracção de Imagens, Extracção de Características e Detecção de Cortes já foram abordados anteriormente nas secções 4.1.2 e 4.1.3, de seguida serão feitas apenas considerações relativas aos métodos de cálculo de similaridade entre imagens par a par, cálculo de valor de semelhança entre cenas e avaliação de semelhanças entre cenas.

4.3.2 Similaridade entre Imagens Par a Par

As similaridades entre imagens de um excerto de vídeo podem ser visualizadas através da construção de uma matriz contendo as distâncias entre as imagens do segmento par a par. Um valor baixo da distância entre descritores corresponde a imagens com elevada semelhança. Esta matriz é simétrica, uma vez que a distância entre o descritor da imagem A e B é igual à distância entre a imagem B e A, e tem dimensão igual ao número de imagens do segmento de vídeo. A distância entre a imagem A e a imagem A é obviamente zero, ou seja a imagem tem similaridade máxima consigo própria.

A Figura 4.3 dá o exemplo de um segmento de vídeo com 51 cenas distintas contendo 6000 imagens (*frames*), totalizando 4 minutos de vídeo. A imagem ilustra uma matriz de similaridade de dimensão 6000x6000, onde cada ponto representa a distância de similaridade entre duas imagens do segmento vídeo utilizando o descritor *Scalable Color*. A representação utiliza valores próximos de zero (preto) para imagens com distâncias de similaridade baixas (imagens similares) e valores próximos de 255 (branco) para imagens com distâncias de similaridade elevadas (imagens não similares).

Na matriz de similaridade, uma área de cor uniforme indica um valor constante de similaridade entre duas sequências de *frames*. Podem também identificar-se diversas

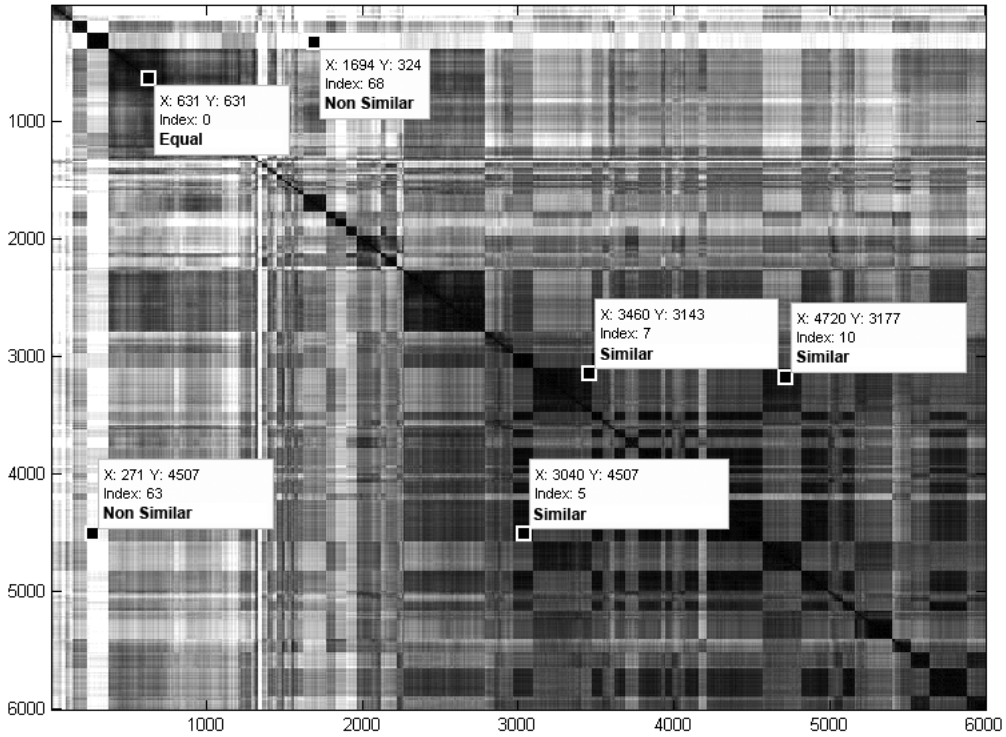


Figura 4.3: Matriz de similaridade de imagens par a par para o descritor *Scalable Color*

regiões, com valores de similaridade uniformes, sendo que as regiões rectangulares a preto representam segmentos de vídeo semelhantes e as regiões claras segmentos de vídeo dissimilares. De notar que as transições abruptas entre regiões de valores de semelhança indicam possíveis cortes de cena.

A Figura 4.3 mostra também 6 pontos exemplificativos de imagens iguais (distância de similaridade 0), imagens similares (distância de similaridade inferior a 10) e imagens dissimilares (distância de similaridade superiores a 10). Duas situações podem ainda ocorrer no caso da presença de imagens iguais ou similares: imagens que pertencem à mesma cena ou imagens que pertencem a cenas distintas. Quando duas imagens têm uma distância de similaridade baixa (imagens iguais ou similares) e pertencem a cenas diferentes, conclui-se que essas cenas são similares e consequentemente a sua metainformação e anotações são candidatas a ser partilhada. A Figura 4.4 mostra as imagens e valores para os 6 pontos referidos. De relembrar que o descritor utilizado é o *Scalable Color*.

Figura 4.4: Exemplos de Similaridades de Cenas

4.3.3 Cálculo Valor de Semelhança entre Cenas

Sendo uma cena constituída por uma série de imagens e tendo em conta que a matriz de similaridades de imagens par a par fornece valores de similaridades entre cada uma das imagens constituintes do excerto de vídeo, só poderão ser efectuadas detecções de cenas similares criando o conceito de cena na matriz de similaridades. É necessário agrupar todas as similaridades par a par correspondentes às imagens constituintes da cena por forma a obter um valores de semelhança de cenas par a par com uma valor representativo para cada um dos pares de cenas. A imagem 1a) da Figura 4.5 mostra as zonas relativas às similaridades das imagens par a par, segundo o descritor *Color*

Layout, para as 4 cenas finais do segmento vídeo *Other Side Of Heaven* (IMDB, 2001). A imagem 1b) mostra um pormenor das similaridades par a par para a zona assinalada a vermelho na imagem 1a). Embora na imagem 1a) as zonas pareçam ter valores de similaridade iguais, na realidade elas variam entre pares de imagens pertencentes a cenas iguais. É necessário encontrar apenas um valor de similaridade que seja representativo da similaridade entre os vários pares de cenas.

As imagens 2a), 2b) e 2c) da Figura 4.5 ilustram o processo de cálculo de valor de semelhança entre pares de cenas.

Partindo da matriz de similaridades de imagens par a par (Imagem 1a da Figura 4.5) e dos valores obtidos pela detecção de cortes de cena, são criadas regiões que incluem as similaridades par a par das imagens correspondentes às cenas. Os valores de similaridade nessa região diferem de par de imagens para par de imagens e é neste ponto que se torna necessária a condensação desses vários valores para um único representante da similaridade do par de cenas. A título de exemplo apenas foram consideradas 4 cenas distintas resultando num conjunto de 16 pares de cenas.

De seguida, para cada uma dessas regiões é determinado o valor Máximo, Médio e Mínimo das similaridades das imagens par a par na região (imagens 2a, 2b e 2c da Figura 4.5). Na Figura 4.5 (imagens 2a, 2b e 2c) podem observar-se as 16 regiões delimitadas pelos cortes de cena e respectivos valores de similaridade com os seus pares. Tal como na matriz de similaridade de imagens par a par, estas matrizes são simétricas e na diagonal apresentam os valores de similaridades de cenas com elas próprias. Para a diagonal no caso de ser considerado o valor Mínimo na região a similaridade toma o valor 0 (similaridade máxima). No caso dos valores Máximo e Média, as regiões da diagonal podem não apresentar valores de similaridade com valor 0 (Máxima). Isto acontece se houver pares de imagens da cena com valores elevados de dissemelhança, o que normalmente ocorre quando a cena é constituída por imagens não uniformes ou apresenta muito movimento.

Deste processo resultam, para cada par de cenas, os 3 valores representativos da similaridade entre cenas. Na detecção de cenas similares poderão ser utilizados individualmente os valores de similaridade entre duas cenas derivados da matriz de similaridades de imagens par a par, ou combinações entre estes valores por forma a maximizar a eficácia de detecção de cenas similares. No processo de avaliação de semelhança entre cenas, descrito na secção seguinte, serão utilizados individualmente os valores de Máximo, Mínimo e Média. A análise de resultados experimentais, que se segue, apresenta

os resultados comparativos destas métricas.

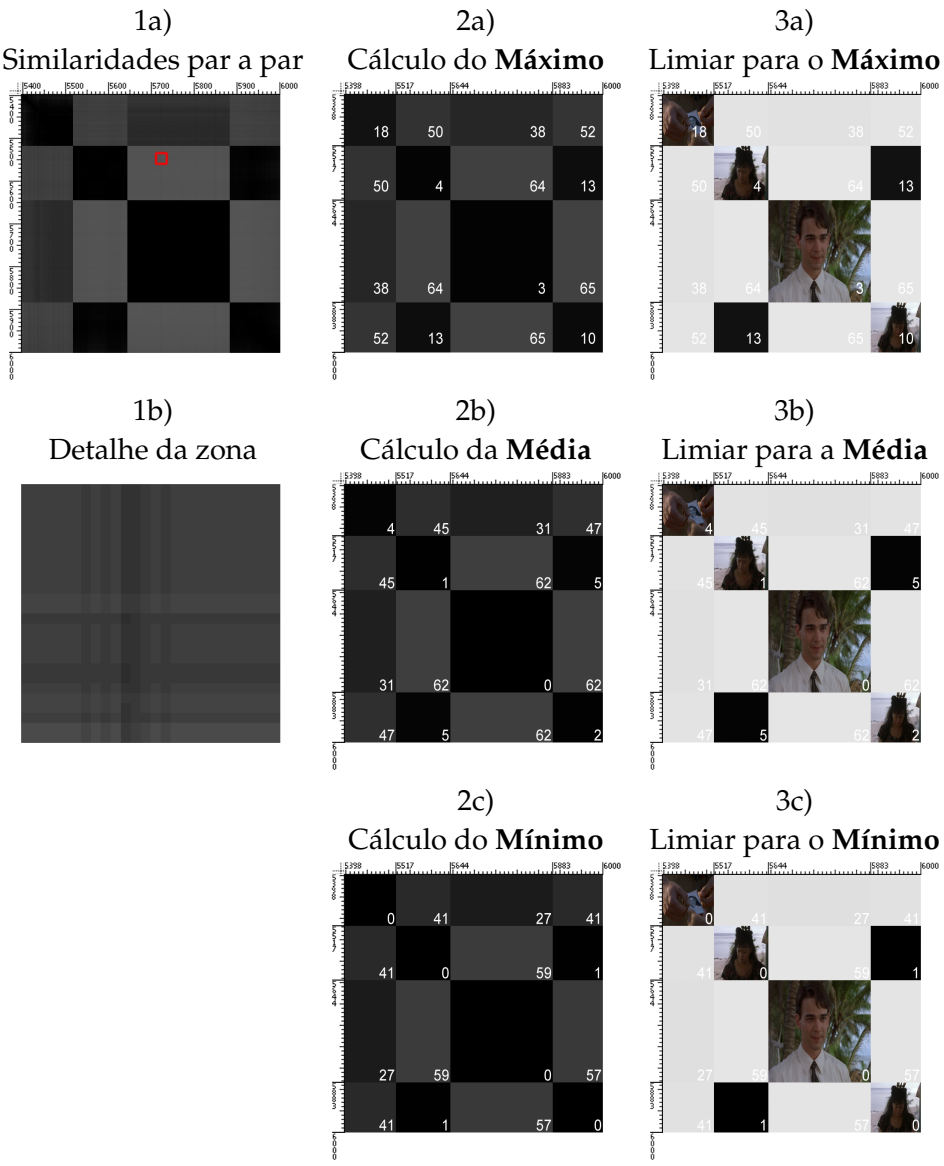


Figura 4.5: Exemplificação do processo de detecção e correspondência de cenas similares

4.3.4 Avaliação de Semelhança entre Cenas

Para fazer a avaliação de semelhança das cenas é elaborada, para cada uma das cenas do segmento, uma lista ordenada por valor de similaridades das cenas.

Para obter as cenas com pares similares, é aplicado a cada uma das matrizes de Máxi-

mo, Média e Mínimo um valor de Limiar que identifica pares de cenas similares acima de um determinando valor de semelhança (Imagens 3a, 3b e 3c da Figura 4.5). No exemplo foi aplicado um limiar valor de semelhança 20, de onde resultou a detecção de um par de cenas similares: cenas delimitadas pelos cortes de cena 5517 a 5644 e 5883 a 6000, com valores de semelhança 13 para o Máximo (Imagem 3a da Figura 4.5), 5 para a Média (Imagem 3b da Figura 4.5) e 1 para o Mínimo (Imagem 3c da Figura 4.5).

A Figura 4.6 ilustra não só as cenas que estão acima do valor de similaridade (pares em cor normal), como também efectua uma ordenação por similaridade independentemente do valor de limiar. As cenas ilustradas em fundo esbatido são aquelas abaixo do limiar de semelhança.

Figura 4.6: Ordenação de imagens por similaridade usando o descritor *Scalable Color* e a métrica de Máximo

A identificação de cenas similares com recurso a um valor de limiar é muito dependente do tipo de cenas e do valor de limiar em si. A escolha de um valor limiar de similaridade que seja eficaz em todas as situações é tarefa difícil senão mesmo impossível. Um valor de limiar alto faz com que sejam retornadas poucas cenas similares. Quando o limiar é baixo, retorna demasiadas cenas. Uma forma encontrada para contornar este problema é o de considerar apenas as n cenas que tenham o valor de similaridade mais alta (n cenas mais similares). Neste caso a abordagem peca por excesso uma vez que

retorna sempre n cenas similares mesmo que não haja efectivamente nenhuma similaridade entre cenas. Neste caso pode ainda aplicar-se um valor de limiar à lista das n cenas similares, cortando o tamanho da resposta a cenas efectivamente dentro de um limiar, eventualmente menos estrito.

Os resultados de similaridades de cenas (avaliado pelos valores de recuperação e precisão) foram calculados utilizando os 3 métodos distintos de avaliação de similaridade descritos:

- 5 cenas mais semelhantes
- 5 cenas mais semelhantes acima do limiar de semelhança
- cenas semelhantes acima do limiar de semelhança

Como referido, para cada um destes métodos de avaliação, são ainda consideradas as 3 métricas de similaridade da cena, Máximo, Média e Mínimo, e cada um dos Descritores *Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Scalable Color*. Na secção seguinte são apresentados resultados obtidos num ambiente experimental para cada uma das combinações de detecção de cenas similares descritas.

4.4 Resultados Experimentais

Para obtenção de resultados experimentais foram utilizados 5 segmentos de vídeo de 4 minutos cada totalizando 20 minutos (aproximadamente 30.000 imagens). Os segmentos vídeo escolhidos diferem no conteúdo tendo sido escolhidos excertos de um vídeo de testes do MPEG-7 (*Animal*) (MPEG, 1998), de um noticiário (*Noticias TVE*) (MPEG, 1998), de um concurso televisivo (*Concurso TVE*) (MPEG, 1998), de um filme de animação (*Inspector Gadget*) (imdb, 2005) e de um filme (*Other Side Of Heaven*) (IMDB, 2001). Por forma a simplificar o problema, devido ao tamanho do conjunto de teste, foram apenas calculadas semelhanças de cenas dentro do próprio excerto vídeo. Diminuiu-se desta forma o tamanho das matrizes de semelhança de 30.000 x 30.000 valores de semelhança para matrizes de 6.000 x 6.000[1].

[1]De referir que a ordem de grandeza destas matrizes de similaridades é comparável à envolvida no processamento de imagens de 36 *Mega-pixels*.

A análise de resultados obtidos incidiu tanto na detecção de cortes de cena como na identificação de cenas similares entre segmentos. Em conjunto, estes dois processos constituem a base para a identificação de cenas com características similares e consequentemente com elevada possibilidade de reutilização de metainformação. As subsecções seguintes apresentam esses resultados agrupados por excerto de vídeo analisado. No caso de similaridade entre cenas são apresentados os resultados obtidos por cada um dos descritores *Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Scalable Color*.

4.4.1 Resultados de Detecção de Cortes de Cena

Segundo os cenários de utilização apresentados na Secção 3.5, a reutilização de anotações traz vantagens na anotação de variações de conteúdo. Para a detecção de cenas similares em vídeos iguais, tendo por exemplo resoluções diferentes, o detector de cortes de cena deverá ser suficientemente robusto para detectar igual número de cortes, nas mesmas posições, independentemente da resolução. A avaliação da independência do detector de cortes de cena em resoluções espaciais diferentes (resolução normal e metade da resolução) foi feita no segmento video *Other Side Of Heaven* (IMDB, 2001) através do descritor *Video Editing*.

	Total de Cortes	Detectados	Não Detectados	Vizinhos	Falsos Positivos
Resolução Normal	100% 50 cortes	86% 43 cortes	14% 7 cortes	10% 5 cortes	231 cortes
Metade da Resolução	100% 50 cortes	78% 39 cortes	22% 11 cortes	16% 8 cortes	230 cortes

Tabela 4.12: Avaliação da imunidade do descritor *Video Editing* a variações de resolução espacial

A Tabela 4.12 mostra os resultados de cortes obtidos para ambas as resoluções comparadas com uma lista de cortes obtida por inspecção visual. Também mostra os valores de eficácia do descritor *Video Editing* para ambas as resoluções. Da análise de ambas pode concluir-se que o descritor é bastante imune a variações de resolução espacial. Há uma diferença de 4 cortes de cena não detectados na resolução mais baixa, mas por

outro lado pode verificar-se que a detecção desses cortes foi feita numa das imagens vizinhas[2]. O valor de eficácia dos cortes de cena da Tabela 4.12 ilustra esse facto: 96% de cortes de cena detectados ou vizinhos para a resolução normal e 94% de cortes de cena detectados ou vizinhos para metade da resolução. O descritor revela no entanto uma elevada taxa de falsos positivos para ambos os casos (231 e 230 respectivamente). Aqui uma cena é dividida quando efectivamente não ocorreu um corte. Com auxílio das similaridades par a par poderá com relativa facilidade ser implementado um algoritmo que verifique quais desses falsos positivos podem ser eliminados ou efectivamente representam um corte de cena. Para isso bastará verificar o valor de semelhança obtido para as imagens imediatamente anterior e posterior. Se o valor de semelhança for muito elevado (corte de cena entre imagens muito semelhantes) podemos eliminar este corte de cena.

A tabela 4.13 mostra os valores de cortes de cena e eficácia de detecção de cortes de cena obtidos para os restantes segmentos vídeo (*Animal* (MPEG, 1998), *Concurso TVE* (MPEG, 1998), *Inspector Gadget* (imdb, 2005)) e *Noticias TVE* (MPEG, 1998). A taxa média de eficácia de detecção de cortes de cena obtida foi de 75% a 87% consoante sejam considerados cortes detectados na vizinhança ou não.

4.4.2 Resultados de Similaridades de Cenas

Os resultados de similaridade entre cenas estão intrinsecamente ligados à semelhança entre os descritores obtidos para os pares de imagens que as compõem. Da comparação das imagens que compõem o segmento vídeo resultam as matrizes simétricas de similaridade para os vários descritores (à semelhança do processo descrito na Secção 4.3.2). Para cada segmento vídeo são, assim, obtidas 4 matrizes de similaridade, uma por descritor (*Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Scalable Color*), com dimensão *n*x*n*, onde *n* é o número de imagens constituintes do excerto. As Figuras 4.7 a 4.11[3], mostram as matrizes agrupadas por excerto vídeo e descritor.

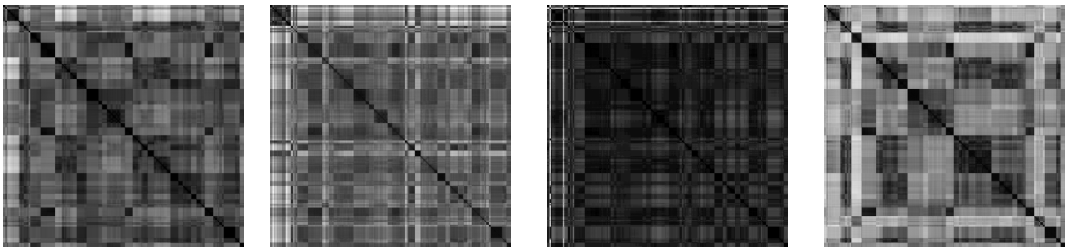
[2]Foi considerado o intervalo entre as 3 imagens anteriores e 3 imagens posteriores.
[3]No Anexo A são apresentadas todas estas matrizes com maior resolução.

	Total de Cortes	Detectados	Não Detectados	Vizinhos	Falsos Positivos
Animals	100%	57%	32%	11%	
	37 cortes	21 cortes	16 cortes	4 cortes	255 cortes
Concurso TVE	100%	100%	0%	0%	
	39 cortes	39 cortes	0 cortes	0 cortes	272 cortes
Inspector Gadget	100%	73%	27%	5%	
	55 cortes	40 cortes	15 cortes	3 cortes	162 cortes
Noticias TVE	100%	57%	43%	26%	
	46 cortes	26 cortes	20 cortes	12 cortes	238 cortes
Other Side Of Heaven	100%	86%	14%	10%	
	50 cortes	43 cortes	7 cortes	5 cortes	231 cortes
Other Side Of Heaven L	100%	78%	22%	16%	
	50 cortes	39 cortes	11 cortes	8 cortes	230 cortes
Total	100%	75%	25%	12%	
	277 cortes	208 cortes	69 cortes	32 cortes	1388 cortes

Tabela 4.13: Resultados de cortes de cena

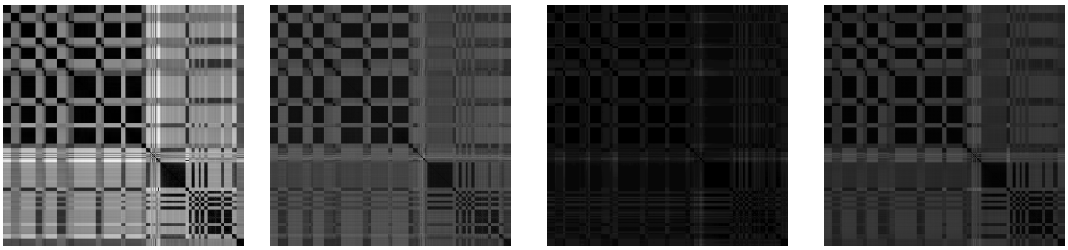
O processo de detecção de similaridade entre cenas tem um custo não negligenciável uma vez que requer a avaliação de todas os pares de similaridades relativas às imagens que constituem a cena. Aplicando a simplificação do problema de identificação de cenas similares descrito na Secção 4.3.1, usando as 3 métricas de uniformização valores de semelhança entre imagens de pares de cenas (valor Máximo, Médio e Mínimo do valor das similaridades entre pares de imagens relativas a um par de cenas), resultam para cada um dos excertos de vídeo 12 matrizes com tamanho nxn, sendo n o número total de cenas[4], são calculadas 3 matrizes (Máximo, Média e Mínimo) para cada um dos 4 Descritores já referidos. Para cálculo dos resultados de similaridade entre cenas consideraram-se somente as similaridades relativas a pares de cenas pertencentes ao mesmo excerto vídeo. O número de cenas consideradas para cálculo de similaridade

[4]Por forma a tornar independentes a análise de resultados obtidos pelo algoritmo de detecção de cortes de cena e a detecção de similaridades de cenas, foram considerados os cortes de cena obtidos por inspecção visual e não os referidos na Secção 4.4.1.



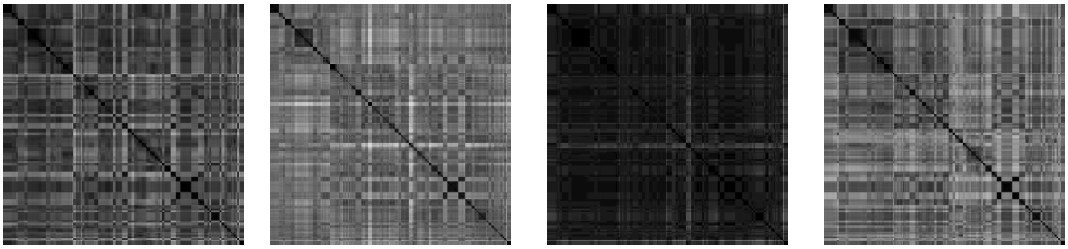
Color Layout *Edge Histogram* *Homogeneous Texture* *Scalable Color*

Figura 4.7: Matrizes de similaridade para o excerto de vídeo *Animals*



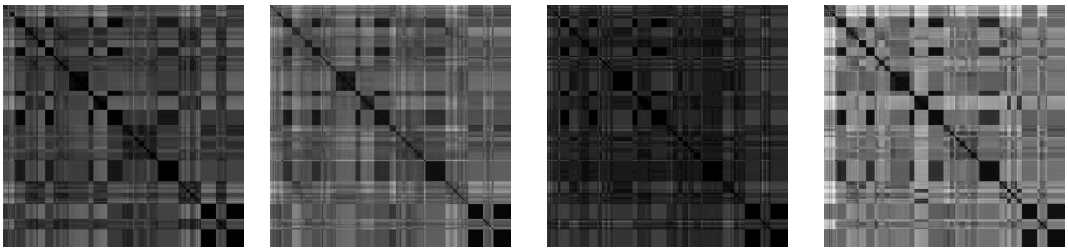
Color Layout *Edge Histogram* *Homogeneous Texture* *Scalable Color*

Figura 4.8: Matrizes de similaridade para o excerto de vídeo *Concurso TVE*



Color Layout *Edge Histogram* *Homogeneous Texture* *Scalable Color*

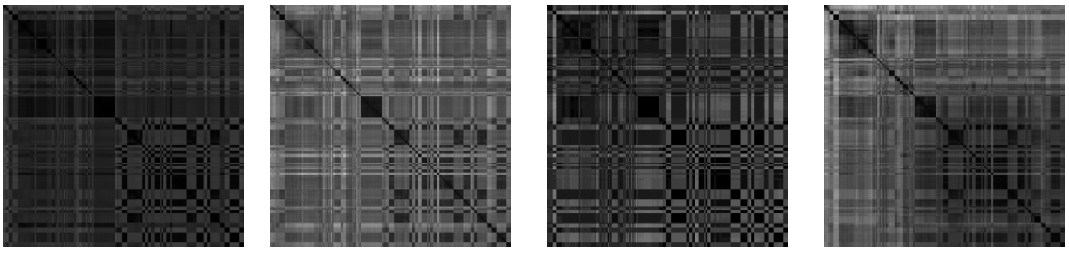
Figura 4.9: Matrizes de similaridade para o excerto de vídeo *Inspector Gadget*



Color Layout *Edge Histogram* *Homogeneous Texture* *Scalable Color*

Figura 4.10: Matrizes de similaridade para o excerto de vídeo *Noticias TVE*(6000x6000)

foi de 37 cenas para o excerto vídeo *Animals*, 39 para o *Concurso TVE*, 55 para o *Inspector Gadget*, 46 para o *Noticias TVE* e 50 cenas para o *Other Side Of Heaven* (IMDB, 2001), tal como mostrado na Tabela 4.13 coluna Total de Cortes.



Color Layout *Edge Histogram* *Homogeneous Texture* *Scalable Color*

Figura 4.11: Matrizes de similaridade para o excerto de vídeo *Other Side Of Heaven*

Uma vez que os resultados obtidos são ilustrados com um número elevado de gráficos comparativos optou-se por incluir esses gráficos em anexo.

No Apêndice C encontram-se, todos os gráficos de resultados obtidos na análise de similaridade entre cenas. Estes resultados compreendem os valores de recuperação e de precisão agrupados em intervalos de 20 pontos percentuais e o seu objectivo é o de perceber qual é o descritor, ou descritores, e em que condições fornecem a percentagem mais elevada de similaridades detectadas com taxa de recuperação e precisão acima dos 80%. Seguindo duas abordagens distintas de análise comparativa de resultados, no Apêndice C.1 ilustram-se resultados comparativos do desempenho dos vários descritores, considerando independentemente a análise para Máximos, Médias e Mínimos de semelhanças e métodos de avaliação de cenas similares já referidos: 5 cenas mais semelhantes, 5 cenas mais semelhantes acima do limiar de semelhança e cenas semelhantes acima do limiar de semelhança. Neste caso foram considerados os valores cumulativos de recuperação e precisão relativas aos vários excertos vídeo.

O Apêndice C.2 apresenta os valores de recuperação e precisão, também em intervalos de 20 pontos percentuais, agrupados por excerto de vídeo e tipo de descritor. Neste caso dá-se ênfase à análise comparativa do desempenho de cada Descritor dependendo do excerto de vídeo em questão. Para cada um dos descritores são mostrados os resultados obtidos pelos Máximos, Médias e Mínimos de semelhanças e métodos de avaliação de cenas similares: 5 cenas mais semelhantes, 5 cenas mais semelhantes acima do limiar de semelhança e cenas semelhantes acima do limiar de semelhança.

A título de exemplo e para facilitar a apresentação dos resultados do Apêndice C, foram elaboradas, para cada um dos conjuntos de gráficos dos Apêndices C.1 e C.2, tabelas que contemplam os melhores valores de percentagem de cenas recuperadas com intervalo de precisão e recuperação acima dos 80%. Por exemplo, na Tabela 4.14 a primeira célula indica que 31% das cenas apresentam uma taxa de recuperação aci-

	Máximo	Média	Mínimo
5 Mais	31% Scalable Color	35% Scalable Color	32% Scalable Color
5 Mais Acima Limiar	17% Scalable Color	18% Color Layout	19% Color Layout
Acima Limiar	19% Color Layout/Scalable Color	36% Color Layout	48% Color Layout

Tabela 4.14: Melhores taxas de recuperação e respectivo descritor (intervalo [80%,100%])

	Máximo	Média	Mínimo
5 Mais	19%+0% Scalable Color	18%+0% Scalable Color	18%+0% Scalable Color
5 Mais Acima Limiar	20%+11% Scalable Color	29%+15% Color Layout	30%+10% Color Layout
Acima Limiar	21%+15% Scalable Color	29%+15% Color Layout	27%+10% Edge Histogram

Tabela 4.15: Melhores taxas de precisão e respectivo descritor (intervalo [80%,100%[e 100%)

ma de 100% para as condições indicadas (5 cenas mais semelhantes, com avaliação do Máximo para o descritor *Scalable Color*).

As Tabelas 4.14 e 4.15 mostram e resumem os gráficos ilustrados no Apêndice C.1, para todo o conjunto das cenas de todos os excertos de vídeo, os valores máximos de número de cenas com taxas de recuperação e precisão acima dos 80% e respectivo Descritor associado.

Relativamente ao Apêndice C.2, as Tabelas 4.16 e 4.17 resumem os melhores resultados obtidos nas gamas de recuperação e precisão acima dos 80% (intervalo [80%,100%]). Associado ao valor de recuperação (Tabela 4.16) encontra-se também ilustrado o método de uniformização de Similaridades das imagens que constituem a cena (Máximo, Média ou Mínimo) e tipo de algoritmo utilizado para avaliação de cenas similares. Nalguns casos o valor máximo das taxas de recuperação e precisão foi obtido com

	Color Layout	Edge Histogram	Homogeneous Texture	Scalable Color
Animals	79% Mínimo Acima Limiar	26% Mínimo 5 Mais	18% Média/Mínimo 5 Mais/Acima Limiar	50% Média 5 Mais
Noticias TVE	43% Mínimo Acima Limiar	34% Média 5 Mais	19% Média 5 Mais	32% Média 5 Mais
Concurso TVE	95% Mínimo Acima Limiar	59% Média 5 Mais	79% Média/Média 5 Mais/Acima Limiar	87% Máximo Acima Limiar
Inspector Gadget	16% Média 5 Mais	14% Média 5 Mais	13% Mínimo Acima Limiar	14% Média/Mínimo 5 Mais
Other Side Of Heaven	33% Média/Mínimo Acima Limiar	29% Mínimo Acima Limiar	45% Média/Mínimo 5 Mais	37% Mínimo Acima Limiar

Tabela 4.16: Melhores taxas de recuperação (intervalo [80%,100%])

dois dos métodos, e são apresentados na tabela usando a notação: Média/Mínimo e 5 mais/acima limiar. Os valores de precisão encontram-se ilustrados na Tabela 4.17; aqui, além dos valores de precisão no intervalo [80%,100%[são também apresentados os valores de taxas de precisão de 100% (parte direita do sinal mais).

4.5 Análise de Resultados

A elaboração de um Ambiente Experimental, descrito na Secção 4.3, e obtenção dos respectivos Resultados Experimentais referidos na Secção 4.4 foi a forma escolhida para efectuar a prova de conceito do sistema de Anotação Baseada em Pesquisa.

O sistema de Anotação Baseado em Pesquisa descrito assenta nos processos de recuperação baseada em conteúdo (imagens vídeo) e baseada em anotação (texto). Sendo que a recuperação de texto tem sido alvo de grandes progressos nos últimos anos, no

	Color Layout	Edge Histogram	Homogeneous Texture	Scalable Color
Animals	11%+3% Máximo 5 Mais Acima Limiar	21%+3% Mínimo/Mínimo 5 Mais Acima Limiar/Acima Limiar	11%+5% Máximo/Média 5 Mais Acima Limiar/Acima Limiar	16%+11% Máximo/Média 5 Mais Acima Limiar/Acima Limiar
Noticias TVE	21%+6% Média 5 Mais Acima Limiar	17%+19% Média/Média 5 Mais Acima Limiar/Acima Limiar	13%+19% Média/Média 5 Mais Acima Limiar/Acima Limiar	17%+19% Média/Média 5 Mais Acima Limiar/Acima Limiar
Concurso TVE	72%+0% Máximo/Máximo 5 Mais Acima Limiar/Acima Limiar	77%+0% Média 5 Mais Acima Limiar	74%+0% Média 5 Mais Acima Limiar	79%+0% Média 5 Mais Acima Limiar
Inspector Gadget	18%+13% Média/Média 5 Mais Acima Limiar/Acima Limiar	11%+21% Média/Média 5 Mais Acima Limiar/Acima Limiar	11%+21% Média/Média 5 Mais Acima Limiar/Acima Limiar	20%+18% Mínimo/Mínimo 5 Mais Acima Limiar/Acima Limiar
Other Side Of Heaven	41%+2% Mínimo Acima Limiar	45%+16% Mínimo/Mínimo 5 Mais Acima Limiar/Acima Limiar	33%+16% Média/Média 5 Mais Acima Limiar/Acima Limiar	31%+18% Média/Média 5 Mais Acima Limiar/Acima Limiar

Tabela 4.17: Melhores taxas de precisão (intervalo [80%,100%[e 100%)

ambiente experimental a ênfase foi na vertente de recuperação baseada em conteúdo, cujo objectivo primordial é o da detecção de cenas similares no próprio excerto vídeo ou outros excertos. A identificação de cenas similares servirá de base à reutilização de anotações no contexto de um processo de anotação interactivo assistido por computador. É assim de crucial importância que seja feita a detecção de cenas similares, com recurso a descritores de conteúdo, com taxas de recuperação aceitáveis (fixada aqui acima dos 80%) permitindo uma boa taxa de reutilização de anotações.

Para uma boa recuperação de cenas similares é preciso não só garantir que se consegue identificar de forma eficiente as diferentes cenas que constituem um segmento, mas também garantir que através da utilização de descritores de conteúdo e algoritmos de identificação de semelhança adequados se identificam cenas semelhantes passíveis de reutilização das suas anotações.

No que respeita à identificação das diferentes cenas constituintes de um excerto vídeo, dos resultados apresentados na tabela 4.13, podemos considerar que o algoritmos de detecção de corte de cenas baseado no descritor *Video Editing* oferece um bom desempenho. A taxa de detecção de cortes máxima é de 100% para o excerto vídeo *Concurso TVE*, e o mínimo é de 57% para os excertos *Animals* e *Noticias TVE*. O valor médio obtido para a identificação dos 277 cortes de cena dos vários segmentos situa-se nos 75%, ou seja 208 dos 277 cortes de cena foram detectados com sucesso. Se considerarmos que o corte detectado pelo *Video Editing* se encontra numa das imagens vizinhas (3 imagens anteriores e 3 imagens posteriores) a percentagem de cortes de cena detectados sobe para os 87%, ou seja 240 de 277 cortes de cena. De referir ainda uma boa imunidade do detector de cortes de cena a variações de resolução dos excertos vídeo, observado por exemplo quando se utiliza uma versão de qualidade reduzida (com metade da resolução espacial) para o excerto vídeo *Other Side Of Heaven*, em que apenas não foram detectados 4 cortes de cena relativamente à versão de resolução total. O maior problema do algoritmo de detecção de cortes é a elevada taxa de falsos positivos na identificação de cortes (1388 cortes). Este facto pode ter origem na utilização de excertos de vídeo que já sofrerem alteração de conteúdo por utilização de esquemas de compressão (MPEG-2 no caso dos excertos considerados). Mesmo assim este é um cenário mais próximo da realidade e, uma vez que no âmbito da identificação de similaridades de cenas são construídas matrizes de semelhança, é possível com relativa facilidade eliminar falsos positivos do algoritmo de cortes de cena com recurso à comparação da similaridade par a par entre as duas imagens, anterior e actual. Não é imperativa a utilização deste algoritmo de detecção de cortes, uma vez que com a utili-

zação de detectores de cortes mais complexos podem com certeza ser obtidos melhores resultados.

Na identificação de cenas similares no conjunto de todas as cenas que constituem os vários excertos vídeo, cujos resultados são apresentados nas tabelas 4.14 e 4.15 , foi obtida uma percentagem máxima de 48% das cenas similares com valor de recuperação acima dos 80% utilizando o descritor *Color Layout*, o Mínimo valor das similaridades que constituem a cena, e considerando cenas semelhantes acima do limiar de semelhança. O valor de percentagem mais baixo (17%), foi obtido com o descritor *Scalable Color*, Máximo valor das similaridades que constituem a cena, e recuperando as 5 cenas mais semelhantes acima do limiar de semelhança. Ou seja os resultados referem que 109 das cenas (48% das 227) no máximo, e 39 das cenas (17% das 227) obtiveram uma taxa de recuperação superior ou igual a 80%, nas condições descritas. São os descritores *Color Layout* e *Scalable Color* que distribuem os melhores resultados de identificação de cenas similares, não é conclusiva qual das métricas (Média ou Mínimo), nem algoritmo de identificação (5 cenas mais semelhantes ou cenas semelhantes acima do limiar de semelhança) tem a melhor performance. Combinações entre ambas proporcionam boas percentagens de cenas identificadas com recuperação acima dos 80%. Os valores de precisão obtidos referem percentagens de 29% das cenas com precisão no intervalo acima de 80% e 100% exclusive, e precisão 100% em 15% das cenas para o descritor *Color Layout*, Média do valor das similaridades que constituem a cena, e ambos algoritmos 5 cenas mais semelhantes acima do limiar de semelhança e cenas mais semelhantes acima do limiar de semelhança. Os piores resultados de precisão (18% acima de 80% e 0% com precisão 100%) foram obtidos com o descritor *Scalable Color*, Média ou Mínimo valor das similaridades que constituem a cena, para as 5 cenas mais semelhantes acima do limiar de semelhança. No geral, considerando o compromisso entre os valores de precisão e recuperação, o descritor *Color Layout*, utilizando a Média dos valores das similaridades que constituem a cena, com a recuperação de todas as cenas semelhantes acima do limiar de semelhança, obtém um bom resultado com valores de 36% de cenas identificadas como similares com valor de recuperação acima de 80%, 29% de cenas com precisão acima dos 80% e 15% com precisão a 100%.

Fazendo a análise por excerto de vídeo, resultados apresentados nas tabelas 4.16 e 4.17, o melhor resultado de recuperação foi obtido para o excerto *Concurso TVE*: 95% das cenas foram identificadas com recuperação superior a 80%, utilizando o descritor *Color Layout*, o Mínimo valor das similaridades que constituem a cena e recuperação de cenas semelhantes acima do limiar de semelhança. O pior desempenho foi obtido para

o excerto *Inspector Gadget* com apenas 13% das cenas identificadas com recuperação acima dos 80%, conseguido através do descritor *Homogeneous Texture*, o Mínimo valor das similaridades que constituem a cena e com recuperação das cenas semelhantes acima do limiar de semelhança. Neste caso em relação à precisão o valor melhor foi obtido também para o excerto *Concurso TVE* com 79% das cenas identificadas com precisão superior a 80% (nenhuma obteve precisão a 100%), utilizando o descritor *Scalable Layout*, a Média dos valores das similaridades que constituem a cena e com recuperação de cenas semelhantes acima do limiar de semelhança. O valor mais baixo foi de 11% de cenas com precisão acima dos 80% e 3% das cenas com precisão igual a 100%, para o excerto vídeo *Animals*, descritor *Color Layout*, o Máximo valor das similaridades que constituem a cena e recuperação das 5 cenas mais semelhantes acima do limiar de semelhança.

A análise por excerto de vídeo e descritor permite identificar que há grandes variações dependendo do tipo de conteúdo do excerto vídeo em questão e também algumas variações relativamente ao descritor utilizado. O excerto vídeo *Concurso TVE* é o que apresenta melhores valores identificação de cenas similares com boas taxas de recuperação; efectivamente este excerto vídeo contém bastantes repetições de cenas onde são alternadas várias vezes concorrentes e apresentador, e é um ambiente homogéneo uma vez que são cenas sem planos de *zoom*, *pan*, ou outros. Os resultados piores são obtidos para o excerto vídeo *Inspector Gadget*: além de haver uma pequena percentagem de cenas similares potencialmente identificáveis, o excerto apresenta muitas variações de cor, *zoom*, *pan* nas várias cenas. Talvez por se tratar de um excerto vídeo de animação, ou seja imagem sintetizada, os descritores retornem valores menos precisos de similaridade entre imagens. No que respeita ao tipo de descritor utilizado tanto o *Color Layout* como o *Scalable Color* apresentam bons valores de identificação de cenas com recuperação acima dos 80%. Os outros dois descritores por vezes apresentam também bons valores mas nos excertos analisados apresentaram quase sempre valores inferiores aos obtidos com estes descritores.

Por último de referir que no geral à medida que a recuperação aumenta os valores de precisão diminuem. Ou seja obtemos além das cenas efectivamente similares, cenas que foram marcadas como sendo similares mas não o são. No contexto do Sistema de Anotação Baseado em Pesquisa, taxas baixas de precisão não são efectivamente um grande problema, uma vez que o utilizador através da sua interacção poderá fazer uma avaliação final acerca da veracidade da semelhança ente as cenas e decidir se quer reutilizar a respectiva anotação ou não.

De um modo geral pode concluir-se que a utilização de descritores de conteúdo na identificação de cenas similares é efectiva. Nos excertos analisados no pior dos casos conseguiu-se uma identificação de 13% de cenas com taxa de recuperação acima dos 80%. Este valor indica que em 7 das 55 das cenas do excerto *Inspector Gadget* conseguiu-se recuperar entre 80% a 100% das cenas semelhantes. Pode concluir-se que na maioria dos casos pode ser possível a identificação de cenas similares com recurso a descritores de baixo nível, e consequentemente ser possível a reutilização das suas anotações no âmbito de um sistema de anotação assistida.

Capítulo 5

Conclusões

A recuperação de conteúdos de natureza visual é uma tarefa com interesse em muitas áreas de aplicação e para a qual não existe ainda nem uma abordagem estabelecida nem ferramentas genéricas. Este trabalho centra-se na tarefa da anotação de conteúdos vídeo, que é central na cadeia de processamento de um fornecedor de conteúdos e pode, com a generalização da produção de materiais pessoais, revelar-se cada vez mais importante em ferramentas destinadas a público. No trabalho desenvolvido, o levantamento dos sistemas existentes na área de anotação de conteúdos multimédia foi o ponto de partida para a elaboração de uma proposta de evolução desses sistemas considerando também a análise de conteúdo. Desta investigação surgiu a proposta para um sistema misto de anotação que se serve de técnicas de recuperação baseada em conteúdo para pesquisa de conteúdos similares através da comparação de características de baixo nível. Os resultados obtidos permitem auxiliar o utilizador nas tarefas de anotação com sugestões de metainformação relevante ou relacionada com os conteúdos previamente anotados.

A análise dos sistemas de recuperação multimédia existentes ilustra bem a separação existente entre os dois níveis de representação, o nível de descrição com recurso a descritores de conteúdo, e o nível de descrição através de metainformação textual. Ao longo da apresentação das técnicas de recuperação de informação visual ficaram evidenciadas as dificuldades inerentes a ambas abordagens. Na recuperação por humanos de documentos multimédia através de características de baixo nível, as interrogações têm de ser formuladas através de esboços ou imagens exemplo que, depois de extraídas as respectivas características de baixo nível, servem de base à interrogação

no domínio. Por outro lado, na recuperação baseada em metainformação textual, de fácil interpretação por humanos, evidenciaram-se algumas das dificuldades relacionadas com a interpretação automática da metainformação relativa a conceitos de grande valor semântico. Os sistemas da actualidade, propõem formas distintas de diminuir a distância entre os dois níveis de representação e seguem também tendências de evolução distintas. Na área de recuperação textual estão a ser usadas recentemente técnicas alternativas como a semântica latente ou as ontologias como forma de construção de conceitos associados a descrições textuais; no domínio da análise de conteúdos predominam a técnicas de extracção e análise automática de informação utilizando técnicas de aprendizagem, sejam elas no domínio textual ou no domínio dos descritores de conteúdos. O objectivo final de todos é o de diminuir o chamado fosso semântico. As técnicas existentes estão longe de conseguir dar uma resposta eficaz e totalmente automática na descrição de conteúdos que sirvam de base a sistemas de recuperação de conteúdos multimédia. Esta é de facto a razão pela qual os sistemas de anotação multimédia com recurso a anotação manual ainda prevalecem. Para além disso, o fenómeno recente da etiquetagem social na *web*, que é uma anotação manual de conteúdos com palavras-chave, revelou ser uma fonte de informação útil para diminuir o fosso semântico em sistemas de recuperação multimédia.

O sistema proposto de anotação baseado em pesquisa combina num único sistema técnicas de anotação humana e um sistema de recuperação baseado em similaridades de conteúdo. Neste sistema, o utilizador efectua pesquisas combinadas de texto e conteúdo visual, recebendo sugestões para material multimédia relevante ou relacionado, anteriormente anotados. No cenário de uso previsto, em que o utilizador tem a tarefa de anotar um excerto vídeo, as referências às sugestões dadas pelo sistema são avaliadas pelo utilizador que toma decisões sobre a reutilização da metainformação. A elaboração de um ambiente experimental, apenas com recuperação vídeo baseada em conteúdo, permitiu a obtenção de resultados cujas conclusões verificaram a validade da utilização dos descritores propostos para identificação de segmentos semelhantes. Nos resultados obtidos o sistema mostrou-se especialmente eficaz nos casos em que havia repetições de cenas ou variações de conteúdo vídeo, maximizando a reutilização de metainformação de grande valor semântico existente. Para estes casos o sistema permitiu a total reutilização de metainformação. Para os restantes casos o sistema proposto revelou-se como um bom ponto de partida para associação de características de baixo significado semântico a anotações e conceitos de valor semântico elevado. Estas associações podem ser ainda mais enriquecidas com descritores mais complexos que

resultam em detecção de objectos ou faces, que aumentarão o âmbito de reutilização da metainformação.

No ambiente experimental usado para desenvolver o trabalho o MPEG-7 XM foi a ferramenta escolhida para a extracção de características. O trabalho adicional sobre esta ferramenta incluiu ajustes por forma a suportar vídeo, adaptações aos módulos para extrair em simultâneo conjuntos de vários descritores e a elaboração de ferramentas adicionais de análise de resultados de forma gráfica. A tarefa de adequação de alguns dos descritores desenhados para a recuperação baseada em imagem a um sistema de recuperação baseado em vídeo requereu esforço significativo. No ambiente experimental a inclusão de outros conjuntos de descritores é agora uma tarefa simples uma vez que foi mantida toda a modularidade e extensibilidade do MPEG-7 XM. A elaboração de estudos comparativos mais alargados de descritores MPEG-7, a sua efectividade perante diferentes tipos de conteúdos, e a conjugação de vários descritores por forma a maximizar a descoberta de segmentos semelhantes, são algumas propostas de continuação directa do trabalho. Seria ainda possível, além de considerar um segmento vídeo como uma sequência de imagens estáticas, ter em conta propriedades adicionais do vídeo como o movimento ou o som. Uma evolução do ambiente experimental para obtenção de resultados incluindo essas características seria certamente um bom ponto de partida para um sistema mais completo e fiável na identificação de material relevante ou relacionado. No que respeita ao som, os algoritmos de transcrição de discurso, identificação de oradores e identificação musical poderiam ser um fonte de metainformação de alto valor semântico.

Outro aspecto importante a considerar em termos de trabalho futuro é a avaliação e validação do sistema por parte de utilizadores. Daqui poderiam resultar indicadores acerca da validade do modelo em ambiente real e apontar direcções futuras no desenvolvimento de interacção pessoa-computador. Os resultados nesta área seriam úteis na refinação do modelo de sistema no que respeita a interfaces amigáveis e sistemas de apresentação de resultados que minimizassem os tempos de anotação e maximizassem a reutilização de metainformação descritiva. A elaboração de modelos de dados específicos para o processo ou a introdução de melhoramentos a modelos de dados de etiquetas poderia incentivar a interoperabilidade com outros sistemas e consequentemente aumentar as possibilidades de propagação da metainformação.

No imediato, e seguindo a tendência actual de disponibilização de material audiovisual na *web*, seria interessante, num futuro muito próximo, aplicar o conceito de anotação

baseada em pesquisa e pesquisa baseada em conteúdo a um sistema de partilha de conteúdos. Apesar de hoje em dia haver uma panóplia de sistemas de partilha de vídeo, não existem produtos capazes de pesquisar material audiovisual através da análise de conteúdo. Certamente que esta técnica daria uma nova dimensão aos sistemas de anotação multimédia baseados em etiquetagem social. A generalização do uso destes sistemas na *web* poderia contribuir para a diminuição do fosso semântico, uma vez que existe actualmente muita informação de grande valor semântico introduzido por utilizadores com recurso a ferramentas de anotação distribuída.

Referencias

AHMED, N., NATARAJAN, T. & RAO, K.R. (1974). Discrete Cosine Transform. *IEEE Trans. Computers*, **23**, 90–93. 67

ALATA, O., CARIOU, C., RAMANANJARASOA, C. & NAJIM, M. (1998). Classification of Rotated and Scaled Textures using HMHV Spectrum Estimation and the Fourier-Mellin Transform. In *ICIP (1)*, 53–56. 26

ARFKEN, G. (1985). *Scalar or Dot Product. Cap. 1.3, Mathematical Methods for Physicists*. Academic Press, San Diego CA. 11

ASHLEY, J., FLICKNER, M., HAFNER, J.L., LEE, D., NIBLACK, W. & PETKOVIC, D. (1995). The Query By Image Content (QBIC) System. In M.J. Carey & D.A. Schneider, eds., *SIGMOD Conference*, 475, ACM Press. 27, 41

BARGERON, D., GUPTA, A., GRUDIN, J., SANOCKI, E. & LI, F. (2001). Asynchronous Collaboration around Multimedia and its Application to On-Demand Training. In *HICSS '01: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 4*, 4042, IEEE Computer Society, Washington, DC, USA. 48

BEECH, D., MENDELSOHN, N., MALONEY, M. & THOMPSON, H.S. (2004). XML Schema Part 1: Structures Second Edition. W3C recommendation, W3C, http://www.w3.org/TR/2004/REC-xmlschema-1-20041028/. 61

BERTINI, M., BIMBO, A.D. & PALA, P. (2002). Indexing for reuse of TV news shots. *Pattern Recognition*, **35**, 581–591. 3

BILLHARDT, H., BORRAJO, D. & MAOJO, V. (2002). A context vector model for information retrieval. *J. Am. Soc. Inf. Sci. Technol.*, **53**, 236–249. 14

BIMBO, A.D. & PALA, P. (1997). Visual Image Retrieval by Elastic Matching of User Sketches. *IEEE Trans. Pattern Anal. Mach. Intell.*, **19**, 121–132. 23

BIRON, P.V. & MALHOTRA, A. (2004). XML Schema Part 2: Datatypes Second Edition. W3C recommendation, W3C, http://www.w3.org/TR/2004/REC-xmlschema-2-20041028/. 61

BORMANS, J. & HILL, K. (2002). MPEG-21 Overview. ISO/IEC JTC1/SC29/WG11 N5231. 50

BRIGGS, R. (1985). Knowledge representation in sanskrit and artificial intelligence. *AI Magazine*, **6**, 32–39. 32

BRODATZ, P. (1966). *Textures: A Photographic Album for Artists and Designers.*. Dover Pubns, New York. xvii, 25

BURNETT, I.S., PEREIRA, F., DE WALLE, R.V. & KOENEN, R. (2006). *The MPEG-21 Book*. John Wiley & Sons. 50

CABRAL, R.N. (2006). imgSeek - Image Database. [Online; accessed 27-July-2006]. 4

CAID, W.R. & CARLETON, J.L. (1994). Context Vector-Based Text Retrieval. In *Proceedings of the 4th Annual 1994 IEEE Dual-Use Technologies and Applications Conference*, 131 – 138. 14

CARLETON, J., CAID, W.R. & SASSEEN, R.V. (1995). Using CONVECTIS, A Context Vector-Based Indexing System for TREC-4. In *TREC*. 10

CARSON, C., THOMAS, M., BELONGIE, S., HELLERSTEIN, J.M. & MALIK, J. (1999). Blobworld: A System for Region-Based Image Indexing and Retrieval. In D.P. Huijsmans & A.W.M. Smeulders, eds., *VISUAL*, vol. 1614 of *Lecture Notes in Computer Science*, 509–516, Springer. 41

CHAKRABARTI, K., ORTEGA-BINDERBERGER, M., PORKAEW, K. & MEHROTRA, S. (2000). Similar Shape Retrieval in MARS. In *IEEE International Conference on Multimedia and Expo (II)*, 709–712. 22

CHANG, S.F., CHEN, W., MENG, H.J., SUNDARAM, H. & ZHONG, D. (1997a). VideoQ: An Automated Content Based Video Search System Using Visual Cues. In *ACM Multimedia*, 313–324. 27, 41

CHANG, S.F., SMITH, J.R., BEIGI, M. & BENITEZ, A.B. (1997b). Visual Information Retrieval from Large Distributed Online Repositories. *Commun. ACM*, **40**, 63–71. 39

CHEN, H., SCHATZ, B.R., NG, T.D., MARTINEZ, J., KIRCHHOFF, A. & LIN, C. (1996). A Parallel Computing Approach to Creating Engineering Concept Spaces for Semantic Retrieval: The Illinois Digital Library Initiative Project. *IEEE Trans. Pattern Anal. Mach. Intell.*, **18**, 771–782. 39

CHETVERIKOV, D. (1982). Experiments in the Rotation-Invariant Texture Discrimination Using Anisotropy Features. In *Proceedings of the 6th IEEE Congress on Pattern Recognition*, 1071–1073, Munich. 25

CHIAO, Y.C. & ZWEIGENBAUM, P. (2002). Looking for Candidate Translational Equivalents in Specialized, Comparable Corpora. In *COLING*. 11

CHIARIGLIONE, L. (1996). Short MPEG-1 description. ISO/IEC JTC1/SC29/WG11 N. 52

CHIARIGLIONE, L. (2000). Short MPEG-2 description. ISO/IEC JTC1/SC29/WG11 N. 52

CIEPLINSKI, L., KIM, W.Y., OHM, J.R., PICKERING, M. & YAMADA, A. (2001). Multimedia Content Description Interface - Part 3 Visual. ISO/IEC JTC1/SC29/WG11 N4358. 66

CLEVERDON, C.W. (1970). Progress in documentation: Evaluation of information retrieval systems. *Journal of Documentation*, 55–67. 8

CODD, E.F. (1990). *The Relational Model for Database Management, Version 2*. Addison-Wesley. 33

COHEN, F.S., FAN, Z. & PATEL, M. (1991). Classification of Rotated and Scaled Textured Images Using Gaussian Markov Random Field Models. *IEEE Trans. Pattern Anal. Mach. Intell.*, **13**, 192–202. 26

COLOMBO, C. & BIMBO, A.D. (1999). Color-induced image representation and retrieval. *Pattern Recognition*, **32**, 1685–1695. 18

COLOMBO, C., BIMBO, A.D. & PALA, P. (1999). Semantics in Visual Information Retrieval. *IEEE MultiMedia*, **6**, 38–53. 3

CPB (2005). PBCore - Public Broadcasting Metadata Dictionary Project. Http://www.pbcore.org. 51

CSIRO (2005). The Continuous Media Web (CMWeb). [Online; accessed 30-April-2006]. 48

DAUM, B. & MERTEN, U. (2002). *System Architecture with XML*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 32

DAVENPORT, G., SMITH, T.A. & PINCEVER, N. (1991). Cinematic Primitives for Multimedia. *IEEE Comput. Graph. Appl.*, **11**, 67–74. 18

DAVIES, E.R. (2004). *Machine Vision: Theory, Algorithms, Practicalities*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA. 22

DAVIS, L.S. (1981). Polarograms: A new tool for image texture analysis. *Pattern Recognition*, **13**, 219–223. 25

DAVIS, L.S., JOHNS, S. & AGGARWAL, J.K. (1979). Texture Analysis Using Generalized Co-occurrence Matrices. Tech. rep., Austin, TX, USA. 25

DCMI (2000). Dublin Core Metadata Initiative. Http://dublincore.org/. 50, 51

DECKER, S., MELNIK, S., VAN HARMELEN, F., FENSEL, D., KLEIN, M.C.A., BROEKSTRA, J., ERDMANN, M. & HORROCKS, I. (2000). The Semantic Web: The Roles of XML and RDF. *IEEE Internet Computing*, **4**, 63–74. 33

DEL.ICIO.US (2006). del.icio.us - Social Bookmarking. [Online; accessed 12-May-2006]. 47, 55

DOMINICH, S., LALMAS, M. & VAN RIJSBERGEN, C.J. (2000). ACM SIGIR 2000 Workshop on Mathematical/Formal Methods in Information Retrieval. *SIGIR Forum*, **34**, 18–23. 10

DRIMBAREAN, A. & WHELAN, P.F. (2001). Experiments in colour texture analysis. *Pattern Recogn. Lett.*, **22**, 1161–1167. 25

EPFL (2005). Swiss Federal Institute of Technology, COALA (Content-Oriented Audiovisual Library Access)-LogCreator. [Online; accessed 30-April-2006]. 48

EXOO, G. (2003). A Euclidean Ramsey Problem. *Discrete & Computational Geometry*, **29**, 223–227. 21, 72

FAHY, M., FELLER, J., FINNEGAN, P. & MURPHY, C. (2003). Using XML vocabularies to exploit changing business models: the newsML experience. In *ECIS*. 50, 53

FENSEL, D., HORROCKS, I., HARMELEN, F., MCGUINNESS, D. & PATEL-SCHNEIDER, D. (2001). OIL: Ontology Infrastructure to Enable the Semantic Web. 4, 35, 49

FLICKR (2006). Flickr - Photo Sharing. [Online; accessed 12-May-2006]. 4, 47, 55

FOUNTAIN, S.R. & TAN, T. (1998). Efficient Rotation Invariant Texture Features for Content-based Image Retrieval. *Pattern Recognition*, **31**, 1725–1732. 26

FRANKEL, C., SWAIN, M.J. & ATHITSOS, V. (1996). WebSeer: An Image Search Engine for the World Wide Web. Tech. Rep. TR-96-14. 41

FURNAS, G.W., DEERWESTER, S.C., DUMAIS, S.T., LANDAUER, T.K., HARSHMAN, R.A., STREETER, L.A. & LOCHBAUM, K.E. (1988). Information Retrieval using a Singular Value Decomposition Model of Latent Semantic Structure. In Y. Chiaramella, ed., *SIGIR*, 465–480, ACM. 14

GEVERS, T. & SMEULDERS, A.W.M. (1999). The PicToSeek WWW Image Search System. In *ICMCS, Vol. 1*, 264–269. 40

GEVERS, T. & SMEULDERS, A.W.M. (2004). *Content-Based Image Retrieval: An Overview*, chap. 8. IMSC Press Multimedia Series, Prentice Hall, 1st edn. 38, 39

GIFT (2006). GNU Image Finding Tool. [Online; accessed 27-July-2006]. 4

GLUSHKO, R.J., TENENBAUM, J.M. & MELTZER, B. (1999). An XML framework for agent-based E-commerce. *Commun. ACM*, **42**, 106–ff. 33

GODBY, C.J., YOUNG, J.A. & CHILDRESS, E. (2004). A Repository of Metadata Crosswalks. *D-Lib Magazine*, **10**. 51

GONZALEZ, R.C. & WOODS, R.E. (2001). *Digital Image Processing*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 21

GROSKY, W.I., NEO, P. & MEHROTRA, R. (1992). A pictorial index mechanism for model-based matching. *Data Knowl. Eng.*, **8**, 309–327. 23

GRUBER, T. (1992). Ontolingua: A mechanism to support portable ontologies. Tech. rep., Stanford University, Knowledge Systems Laboratory. 32

GRUBER, T.R. (1993). A translation approach to portable ontology specifications. *Knowl. Acquis.*, **5**, 199–220. 32, 33

GUARINO, N. (1998). Formal Ontology and Information Systems. In *Proc. of the International Conference on Formal Ontologies in Information Systems (FOIS'98). Trento (Italy)*, IOS Press, ISBN 0922-6389. 32

GUARINO, N. & WELTY, C.A. (2000). Ontological Analysis of Taxonomic Relationships. In *ER*, 210–224. 4, 33, 34

HAAR, A. (1910). Zur Theorie der orthogonalen Funktionensysteme. *Mathematische Annalen*, 331–371. 21, 72

HAMPAPUR, A., GUPTA, A., HOROWITZ, B., SHU, C.F., FULLER, C., BACH, J.R., GORKANI, M. & JAIN, R. (1997). Virage Video Engine. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 188–198. 27, 41

HARMAN, D. (1993). The First Text REtrieval Conference (TREC-1), Rockville, MD, USA, 4-6 November 1992. *Inf. Process. Manage.*, **29**, 411–414. 42

HEALEY, G. (1998). Using Zernike moments for the illumination and geometry invariant classification of multispectral texture. *IEEE Transactions on Image Processing*, **7**, 196–203. 26

HERZOG, E., CZERNIAK, A. & ADLER, O. (2006). Metacafe - Serving the World Best Videos. [Online; accessed 27-July-2006]. 4

HORI, O. & KANEKO, T. (1999). Results of Spatio-Temporal Region DS Core/Validation Experiment. *ISO/IEC JTC1/SC29/WG11/MPEG99/M5414*, Maui, HI. 29

HORN, R.A. & JOHNSON, C.R. (1990). Norms for Vectors and Matrices. *Ch. 5 in Matrix Analysis*. Cambridge University Press. 21, 26, 72, 74

HU, M.K. (1962). Visual pattern recognition by moment invariants. *Information Theory, IEEE Transactions on*, **8**, 179–187. 22

HUANG, C.L. & HUANG, D.H. (1998). A content-based image retrieval system. *Image Vision Comput.*, **16**, 149–163. 23

HUANG, J., KUMAR, S.R., MITRA, M., ZHU, W.J. & ZABIH, R. (1997). Image Indexing Using Color Correlograms. In *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, 762, IEEE Computer Society, Washington, DC, USA. 20

HURLEY, C., CHEN, S. & KARIM, J. (2006). YouTube - Broadcast Yourself. [Online; accessed 27-July-2006]. 4, 47, 48, 55

IMDB (2001). The Other Side of Heaven. [Online; accessed 28-July-2006]. 84, 87, 88, 91, 1, 23, 38, 42

IMDB (2005). Inspector Gadget's Biggest Caper Ever — Imdb, The Internet Movie Database. [Online; accessed 15-May-2008]. 87, 89, 1, 23, 38, 42

IPTC (1965). International Press Telecommunications Council. [Online; accessed 17-January-2007]. 53, 54

IPTC (2000). News Markup Language. [Online; accessed 17-January-2007]. 50, 53

IPTC (2001). Sports Markup Language. [Online; accessed 17-January-2007]. 54

IRANI, M., ANANDAN, P., BERGEN, J., KUMAR, R. & HSU, S. (1996). Efficient representations of video sequences and their applications. *Signal Processing: Image Communication*, **8**, 327–351. 29

JACOBS, C.E., FINKELSTEIN, A. & SALESIN, D. (1995). Fast multiresolution image querying. In *SIGGRAPH*, 277–286. 41

JEANNIN, S. (1999). MPEG-7 Visual part of eXperimental Model Version 11.0. ISO/IEC JTC1/SC29/WG11 N4632. 72, 73

JEANNIN, S. & DIVAKARAN, A. (2001a). MPEG-7 visual motion descriptors. *Circuits and Systems for Video Technology, IEEE Transactions on*, **11**, 720–724. xvii, 28

JEANNIN, S. & DIVAKARAN, A. (2001b). MPEG-7 visual motion descriptors. *IEEE Trans. Circuits Syst. Video Techn.*, **11**, 720–724. 29

JOACHIMS, T. (1997). A Probabilistic Analysis of the Rocchio Algorithm with TFIDF for Text Categorization. In D.H. Fisher, ed., *ICML*, 143–151, Morgan Kaufmann. 40

JONES, K.S. (1968). Automatic Term Classification and Information Retrieval. In *IFIP Congress (2)*, 1290–1295. 8

KALMAN, R.E. (1960). A New Approach to Linear Filtering and Prediction Problems. *Transactions of the ASME–Journal of Basic Engineering*, **82**, 35–45. 64

KASHYAP, R.L. & KHOTANZAD, A. (1986). A Model-based Method for Rotation Invariant Texture Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **8**, 472–481. 25

KAUPPINEN, H., SEPPÄNEN, T. & PIETIKÄINEN, M. (1995). An Experimental Comparison of Autoregressive and Fourier-Based Descriptors in 2D Shape Classification. *IEEE Trans. Pattern Anal. Mach. Intell.*, **17**, 201–207. 23

KENNEY, J.F. & KEEPING, E.S. (1951). Moments About the Mean. *Ch. 7.3 in Mathematics of Statistics, Part 2, 2nd ed.*. Van Nostrand, Princeton, NJ. 22

KOENEN, R. (2002). Short MPEG-2 description. ISO/IEC JTC1/SC29/WG11 N4668. 29, 52

KORFHAGE, R.R. (1997). *Information Storage and Retrieval*. Wiley Computer Publishing. 9, 11

LANCASTER, F.W. (1968). *Information Retrieval Systems: Characteristics, Testing, and Evaluation*. Wiley, New York. 8

LEE, K.W., YOU, W.S. & KIM, J. (1999). Quantitative Analysis for the Motion. Trajectory Descriptor in MPEG-7. *ISO/IEC JTC1/SC29/WG11/MPEG99/M5400*. 29

LENAT, D.B., GUHA, R.V., PITTMAN, K., PRATT, D. & SHEPHERD, M. (1990). CYC: Toward Programs with Common Sense. *Commun. ACM*, **33**, 30–49. 32

LI, S.Z. (1999). Shape Matching Based on Invariants. In O.M. Omidvar, ed., *Progress in Neural Networks*, vol. 6, Intellect Ltd, London. 23

LIAO, S.X. & PAWLAK, M. (1996). On Image Analysis by Moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, **18**, 254–266. 22

LIN, C.Y., TSENG, B.L., NAPHADE, M.R., NATSEV, A. & SMITH, J.R. (2003). VideoAL: a novel end-to-end MPEG-7 video automatic labeling system. In *ICIP (3)*, 53–56. 48

LIS (2006). MPEG-7 eXperimentation Model (XM), Institute for Integrated Systems - Video Group. [Online; accessed 27-July-2006]. 30

LLOYD, S.P. (1982). Least squares quantization in pcm. *IEEE Transactions on Information Theory*, **28**, 129–136. 21

LU, G. & SAJJANHAR, A. (1999). Region-based shape representation and similarity measure suitable for content-based image retrieval. *Multimedia Syst.*, **7**, 165–174. 22

LUCAS, B.D. & KANADE, T. (1981). An Iterative Image Registration Technique with an Application to Stereo Vision. In P.J. Hayes, ed., *IJCAI*, 674–679, William Kaufmann. 64

MANIAN, V. & VÁSQUEZ, R. (1998). Scaled and rotated texture classification using a class of basis functions. *Pattern Recognition*, **31**, 1937–1948. 26

MANJUNATH, B.S. (2002). *Introduction to MPEG-7, Multimedia Content Description Interface*. B. S. Manjunath, P. Salembier and T. Sikora (editors). 50, 51

MANJUNATH, B.S. & MA, W.Y. (1996). Texture Features for Browsing and Retrieval of Image Data. *IEEE Trans. Pattern Anal. Mach. Intell.*, **18**, 837–842. 25

MANNING, C.D. & SCHTZE, H. (1999). *Foundations of Statistical Natural Language Processing*. The MIT Press. 15, 43

MAO, J. & JAIN, A.K. (1992). Texture classification and segmentation using multiresolution simultaneous autoregressive models. *Pattern Recognition*, **25**, 173–188. 25

MARQUES, O. & FURHT, B. (2002). Content-based visual information retrieval. 37–57. 37, 39, 40

MARR, D. (1982). *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. W. H. Freeman, San Francisco. 29

MARSICO, M.D., CINQUE, L. & LEVIALDI, S. (1997). Indexing pictorial documents by their content: a survey of current techniques. *Image Vision Comput.*, **15**, 119–141. 16

MARTÍNEZ, J.M. (2002). MPEG-7 Overview. ISO/IEC JTC1/SC29/WG11 N4674. xvii, 18, 30, 31, 32, 50, 52

MATHES, A. (2004). Folksonomies - Cooperative Classification and Communication Through Shared Metadata. 46, 49

MEHROTRA, R. & GARY, J.E. (1995). Similar-Shape Retrieval In Shape Data Management. *Computer*, **28**, 57–62. 23

MOG-SOLUTIONS (2007). theScribe. [Online; accessed 07-March-2007]. 48

MOKHTARIAN, F. & BOBER, M. (2003). *Curvature Scale Space Representation: Theory, Applications, and MPEG-7 Standardization*. Kluwer Academic Publishers, Norwell, MA, USA. 74

MOKHTARIAN, F., ABBASI, S. & KITTLER, J. (1996). Robust and Efficient Shape Indexing through Curvature Scale Space. In *BMVC*, British Machine Vision Association. 23

MOOERS, C.N. (1947). The Zator-A Proposal: A Machine for Complete Documentation. *Reprinted with preface as Zator Technical Bulletin No. 65 (1951)*. 7

MPEG (1998). Licensing Agreement for the MPEG-7 Content Set. ISO/IEC JTC1/SC29/WG11 N2466. 87, 89, 1, 23, 38, 42

MSSG (2006). MPEG Software Simulation Group (MSSG). [Online; accessed 27-July-2006]. 63, 79

NACK, F. & LINDSAY, A.T. (1999a). Everything You Wanted to Know About MPEG-7: Part 1. *IEEE MultiMedia*, **6**, 65–77. 8, 50

NACK, F. & LINDSAY, A.T. (1999b). Everything You Wanted to Know About MPEG-7: Part 2. *IEEE MultiMedia*, **6**, 64–73. 8, 50

NETPBM (2003). PPM - Netpbm Color Image Format. [Online; accessed 01-July-2006]. 31, 63, 79

NIBLACK, W., BARBER, R., EQUITZ, W., FLICKNER, M., GLASMAN, E.H., PETKOVIC, D., YANKER, P., FALOUTSOS, C. & TAUBIN, G. (1993). The QBIC Project: Querying Images by Content, Using Color, Texture, and Shape. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 173–187. 16, 22

NIEDERMEIER, U., HEUER, J., HUTTER, A. & STECHELE, W. (2002). MPEG-7 Binary Format for XML Data. In *DCC '02: Proceedings of the Data Compression Conference (DCC '02)*, 467, IEEE Computer Society, Washington, DC, USA. 31, 66

NIST (1992). Text REtrieval Conference (TREC). [Online; accessed 27-January-2007]. 42

NIST (2003). TREC Video Retrieval Evaluation. [Online; accessed 11-January-2007]. 42, 61

NOY, N.F. & MCGUINNESS, D.L. (2001). Ontology Development 101: A Guide to Creating Your First Ontology. Tech. Rep. SMI-2001-0880, Stanford University School of Medicine. 33

NUXEO (2000). Nuxeo: open source ECM - Enterprise Content Management. Http://www.nuxeo.com. 51

OGLE, V. & STONEBRAKER, M. (1995). Chabot: Retrieval from a Relational Database of Images. *IEEE Computer, special issue on Content-Based Image Retrieval Systems*, **V. Gudivada and V. Raghavan. (eds), 28(9)**. 16

OHM, J.R., MAKAI, B. & ZIER, D. (1999). Results of MPEG-7 Core Experiment CT1. ISO/IEC JTC1/SC29/WG11 MPEG98/M4739. 29, 30, 61

OJALA, T., RAUTIAINEN, M., MATINMIKKO, E. & AITTOLA, M. (2001). Semantic Image Retrieval With HSV Correlograms. In *Proceedings of the 12th Scandinavian Conference on Image Analysis*, SCIA, Bergen, Norway. 20

OJALA, T., AITTOLA, M. & MATINMIKKO, E. (2002a). Empirical Evaluation of MPEG-7 XM Color Descriptors in Content-Based Retrieval of Semantic Image Categories. In *ICPR (2)*, 1021–1024. 20

OJALA, T., MAENPAA, T., VIERTOLA, J., KYLLONEN, J. & PIETIKAINEN, M. (2002b). Empirical Evaluation of MPEG-7 Texture Descriptors with A Large-Scale Experiment. In *Proc. 2nd International Workshop on Texture Analysis and Synthesis, Copenhagen, Denmark*, 99–102. 20, 26

OJALA, T., PIETIKÄINEN, M. & MÄENPÄÄ, T. (2002c). Multiresolution Gray-Scale and Rotation Invariant Texture Classification with Local Binary Patterns. *IEEE Trans. Pattern Anal. Mach. Intell.*, **24**, 971–987. 26, 27

PARMAR, M.J. (2005). Review: Distributed Multimedia Database Technologies Supported by MPEG-7 and MPEG-21. *Comput. J.*, **48**, 563–564. 51

PASCASIO, A.A. & TERWILLIGER, P. (2003). The pseudo cosine sequences of a distance-regular graph. 11

PENTLAND, A., PICARD, R.W. & SCLAROFF, S. (1994). Photobook: Tools for Content-Based Manipulation of Image Databases. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 34–47. 41

PERSOON, E. & FU, K.S. (1986). Shape discrimination using Fourier descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, **8**, 388–397. 23

PINKERTON, B. (1994). Finding What People Want: Experiences with the WebCrawler. 4

PLONE (2000). Plone CMS - Open Source Content Management System. Http://www.plone.org. 51

PONCELEON, D.B., SRINIVASAN, S., AMIR, A., PETKOVIC, D. & DIKLIC, D. (1998). Key to Effective Video Retrieval: Effective Cataloging and Browsing. In *ACM Multimedia*, 99–107. 41

RANDEN, T. & HUSØY, J.H. (1999). Filtering for Texture Classification: A Comparative Study. *IEEE Trans. Pattern Anal. Mach. Intell.*, **21**, 291–310. 24, 25

RICOH (2005). The ricoh MovieTool. [Online; accessed 30-April-2006]. 47

ROBERTSON, S.E. (1997). The probability ranking principle in IR. 281–286. 12

ROWE, L.A., BORECZKY, J.S. & EADS, C.A. (1994). Indexes for User Access to Large Video Databases. In *Storage and Retrieval for Image and Video Databases (SPIE)*, 150–161. 17

SACK, H. & WAITELONIS, J. (2006). Integrating Social Tagging and Document Annotation for Content-Based Search in Multimedia Data. In *n Proc. of the 1st Semantic Authoring and Annotation Workshop (SAAW2006)*, Athens (GA), USA. 55

SAFAR, M., SHAHABI, C. & SUN, X. (2000). Image Retrieval by Shape: A Comparative Study. In *IEEE International Conference on Multimedia and Expo (I)*, 141–144. 22

SALEMBIER, P. & SMITH, J.R. (2001). MPEG-7 multimedia description schemes. *IEEE Trans. Circuits Syst. Video Techn.*, **11**, 748–759. 61

SALTON, G. (1970). Automatic Text Analysis. *Science*, 335–343. 8

SALTON, G. (1989). *Automatic Text Processing: The Transformation, Analysis, and Retrieval of Information by Computer*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA. 11

SANTINI, S., JAIN, R. & GUPTA, A. (1999). A User Interface for Emergent Sementics in Image Databases. In R. Meersman, Z. Tari & S.M. Stevens, eds., *DS-8*, vol. 138 of *IFIP Conference Proceedings*, 123–143, Kluwer. 39

SCHEMAWEB (2007). Schema Web Directory. [Online; accessed 27-January-2007]. 34

SCHULZRINNE, H., DIMITROVA, N., SASSE, A., MOON, S.B. & LIENHART, R., eds. (2004). *Proceedings of the 12th ACM International Conference on Multimedia, October 10-16, 2004, New York, NY, USA*, ACM. 61, 117

SCHUTZE, H. (1992). Dimensions of Meaning. In *Proceedings of Supercomputing '92, Minneapolis.*, 787–796. 14

SEBE, N., LEW, M.S., ZHOU, X.S., HUANG, T.S. & BAKKER, E.M. (2003). The State of the Art in Image and Video Retrieval. In E.M. Bakker, T.S. Huang, M.S. Lew, N. Sebe & X.S. Zhou, eds., *CIVR*, vol. 2728 of *Lecture Notes in Computer Science*, 1–8, Springer. 3, 55

SIVIC, J. & ZISSERMAN, A. (2003). Video Google: A text retrieval approach to object matching in videos. In *Proceedings of the International Conference on Computer Vision*, vol. 2, 1470–1477. 4

SMEATON, A.F., OVER, P. & KRAAIJ, W. (2004). TRECVID: Evaluating the Effectiveness of Information Retrieval Tasks on Digital Video. In Schulzrinne *et al.* (2004), 652–655. 42

SMITH, J.R. & CHANG, S.F. (1996). VisualSEEk: A Fully Automated Content-Based Image Query System. In *ACM Multimedia*, 87–98. 41

SMITH, J.R. & CHANG, S.F. (1997). Visually Searching the Web for Content. *IEEE MultiMedia*, **4**, 12–20. 41

SMITH, J.R. & CHANG, S.F. (1999). Integrated Spatial and Feature Image Query. *Multimedia Syst.*, **7**, 129–140. 16

SMITH, J.R. & LUGEON, B. (2000). A Visual Annotation Tool for Multimedia Content Description. In *Proc. SPIE Photonics East, Internet Multimedia Management Systems*. 47

SMPTE (2004a). Material Exchange Format (MXF) Descriptive Metadata Scheme - 1. SMPTE 380M-2004. 48, 50, 53

SMPTE (2004b). The MXF File Format Specification. SMPTE 377M-2004. 48

SMPTE (2007). SMPTE Metadata Dictionary . RP210.10-2007. 53

STEVES, M., RANGANATHAN, M. & MORSE, E. (2001). SMAT: Synchronous Multimedia and Annotation Tool. In *HICSS '01: Proceedings of the 34th Annual Hawaii International Conference on System Sciences (HICSS-34)-Volume 9*, 9025, IEEE Computer Society, Washington, DC, USA. 49

STILL, M. (2005). *The Definitive Guide to ImageMagick (Definitive Guide)*. Apress, Berkely, CA, USA. 31, 64

STOJANOVIC, N. (2005). On the Query Refinement in the Ontology-based Searching for Information. *Inf. Syst.*, **30**, 543–563. 35

SZELISKI, R. (1994). Image Mosaicing for Tele-Reality Applications. In *Proc. of 2nd IEEE Workshop on Applications of Computer Vision, Sarasota*, 44–53, IEEE-Computer Society Press. 29

TAO, Y. & GROSKY, W.I. (2001). Spatial Color Indexing Using Rotation, Translation, and Scale Invariant Anglograms. *Multimedia Tools Appl.*, **15**, 247–268. 20

TAUBIN, G. (1992). *Recognition and Positioning of Rigid Objects Using Algebraic and Moment Invariants*. Ph.D. thesis, Providence, RI, USA. 22

TEH, C.H. & CHIN, R.T. (1988). On Image Analysis by the Methods of Moments. *IEEE Trans. Pattern Anal. Mach. Intell.*, **10**, 496–513. 22, 24, 26

TV-ANYTIME (2003). TV-Anytime Forum. [Online; accessed 01-July-2006]. 50, 54

UNESCO (1977). UNESCO Thesaurus. [Online; accessed 07-March-2007]. 35

USCHOLD, M. & JASPER, R. (1999). A Framework for Understanding and Classifying Ontology Applications. In *Twelfth Workshop on Knowledge Acquisition Modeling and Management KAW 99*. 35

VAN RIJSBERGEN, C.J. (1979). *Information Retrieval*. Butterworths, London, UK. 12

VISE, D. & MALSEED, M. (2006). *The Google Story: Inside the Hottest Business, Media, and Technology Success of Our Time*. Delta. 4

W3C (2007a). OWL Web Ontology Language Guide. [Online; accessed 27-January-2007]. 34

W3C (2007b). RDF - Resource Description Framework. [Online; accessed 27-January-2007]. 34

W3C (2007c). RDF Vocabulary Description Language 1.0: RDF Schema. [Online; accessed 27-January-2007]. 34

WALMSLEY, P. & FALLSIDE, D.C. (2004). XML Schema Part 0: Primer Second Edition. W3C recommendation, W3C, http://www.w3.org/TR/2004/REC-xmlschema-0-20041028/. 61

WEAVER, W. & SHANNON, C. (1949). *The Mathematical Theory of Communication*. University of Illinois Press, Urbana, Illinois, republished in paperback 1963. 7

WEIBEL, S. & LAGOZE, C. (1997). An Element Set to Support Resource Discovery - The State of the Dublin Core: January 1997. *Int. J. on Digital Libraries*, **1**, 176–186. 50, 51

WELLS, N., DEVLIN, B., WILKINSON, J., BEARD, M. & TUDOR, P. (2006). *The MXF Book: An Introduction to the Material EXchange Format*. Focal Press. 48

WIKIPEDIA (2006a). del.icio.us — Wikipedia, The Free Encyclopedia. [Online; accessed 12-May-2006]. 47, 55

WIKIPEDIA (2006b). flickr — Wikipedia, The Free Encyclopedia. [Online; accessed 12-May-2006]. 47, 55

WIKIPEDIA (2006c). Information — Wikipedia, The Free Encyclopedia. [Online; accessed 07-November-2006]. 7

WIKIPEDIA (2006d). Multimedia — Wikipedia, The Free Encyclopedia. [Online; accessed 30-April-2006]. 2

WIKIPEDIA (2007a). Bit-plane — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 67

WIKIPEDIA (2007b). Bitstream — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 28

WIKIPEDIA (2007c). Boolean logic — Wikipedia, The Free Encyclopedia. [Online; accessed 10-January-2007]. xvii, 10

WIKIPEDIA (2007d). Color Histogram — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. xvii, 20

WIKIPEDIA (2007e). Correlation — Wikipedia, The Free Encyclopedia. [Online; accessed 10-January-2007]. 11

WIKIPEDIA (2007f). Covariance — Wikipedia, The Free Encyclopedia. [Online; accessed 10-January-2007]. 11

WIKIPEDIA (2007g). Fuzzy Logic — Wikipedia, The Free Encyclopedia. [Online; accessed 07-January-2007]. 16

WIKIPEDIA (2007h). Fuzzy Set — Wikipedia, The Free Encyclopedia. [Online; accessed 07-January-2007]. 16

WIKIPEDIA (2007i). Fuzzy Set — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. xvii, 16

WIKIPEDIA (2007j). GIF — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 31

WIKIPEDIA (2007k). HSL — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. xvii, 19

WIKIPEDIA (2007l). JPEG — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 31

WIKIPEDIA (2007m). Latent semantic analysis — Wikipedia, The Free Encyclopedia. [Online; accessed 10-January-2007]. xvii, 15

WIKIPEDIA (2007n). Parametric model — Wikipedia, The Free Encyclopedia. [Online; accessed 10-January-2007]. 30

WIKIPEDIA (2007o). Peter Roget — Wikipedia, The Free Encyclopedia. [Online; accessed 07-January-2007]. 35

WIKIPEDIA (2007p). PNG — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 31

WIKIPEDIA (2007q). RGB — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. xvii, 19

WIKIPEDIA (2007r). SQL — Wikipedia, The Free Encyclopedia. [Online; accessed 07-January-2007]. 9

WIKIPEDIA (2007s). Thesaurus — Wikipedia, The Free Encyclopedia. [Online; accessed 07-January-2007]. 4, 35

WIKIPEDIA (2007t). Vector space — Wikipedia, The Free Encyclopedia. [Online; accessed 10-January-2007]. xvii, 10

WIKIPEDIA (2007u). WAV — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 30

WIKIPEDIA (2007v). YCbCr — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 67

WIKIPEDIA (2007w). YouTube — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. 47, 48, 55

WIKIPEDIA (2007x). YUV — Wikipedia, The Free Encyclopedia. [Online; accessed 07-March-2007]. xvii, 19

WITTEN, I.H., MOFFAT, A. & BELL, T.C. (1999). *Managing Gigabytes: Compressing and Indexing Documents and Images*. Morgan Kaufmann. 9, 43

WONG, S.K.M. & RAGHAVAN, V.V. (1984). Vector Space Model of Information Retrieval: A Reevaluation. In *SIGIR '84: Proceedings of the 7th annual international ACM SIGIR conference on Research and development in information retrieval*, 167–185, British Computer Society, Swinton, UK, UK. 10

WONG, S.K.M. & YAO, Y.Y. (1995). On modeling information retrieval with probabilistic inference. *ACM Trans. Inf. Syst.*, **13**, 38–68. 9

YAMAMOTO, D. & NAGAO, K. (2004). iVAS: Web-based Video Annotation System and its Applications. In *International Semantic Web Conference*, 486–501. 48

YANG, J. & FILO, D. (2006). Yahoo! [Online; accessed 27-July-2006]. 4

YOU, J. & COHEN, H.A. (1993). Classification and segmentation of rotated and scaled textured images using texture "tuned"masks. *Pattern Recognition*, **26**, 245–258. 26

ZADEH, L.A. (1965). Fuzzy Sets. *Information and Control*, **8**, 338–353. 15

ZGDV (2005). Zentrum fuer Graphische Datenverarbeitung e.V., VIDETO - Video Description Tool. [Online; accessed 30-April-2006]. 48

ZHAI, G., FOX, G.C., PIERCE, M., WU, W. & BULUT, H. (2005). eSports: Collaborative and Synchronous Video Annotation System in Grid Computing Environment. In *ISM '05: Proceedings of the Seventh IEEE International Symposium on Multimedia*, 95–103, IEEE Computer Society, Washington, DC, USA. 49

ZHANG, D. & LU, G. (2002). A Comparative Study of Fourier Descriptors for Shape Representation and Retriev. *Int. J. Image Graphics*, **2**, 269–285. 23

ZHANG, D. & LU, G. (2003). Evaluation of MPEG-7 Shape Descriptors Against Other Shape Descriptors. *Multimedia Syst.*, **9**, 15–30. xvii, 20, 22

ZHENG, X. & GAO, Q. (2004). Detection of Perceptual Junctions by Curve Partitioning and Grouping. In *CRV '04: Proceedings of the 1st Canadian Conference on Computer and Robot Vision (CRV'04)*, 347–353, IEEE Computer Society, Washington, DC, USA. xvii, 23

ZHOU, T.T. & JIN, J.S. (2004). Principles of Video Annotation Markup Language (VAML). In *VIP '05: Proceedings of the Pan-Sydney area workshop on Visual information processing*, 123–127, Australian Computer Society, Inc., Darlinghurst, Australia, Australia. 50, 54

ZOBEL, J. (1998). How Reliable Are the Results of Large-Scale Information Retrieval Experiments? In *SIGIR*, 307–314, ACM. 42

ZOPE (2005). Zope Community - Open Source Application Server. Http://www.zope.org. 51

Apêndice A

Matrizes de Similaridade

Neste anexo são incluídas as matrizes de similaridade para os descritores *Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Scalable Color* agrupados por excerto de vídeo *Animal* (MPEG, 1998), *Noticias TVE* (MPEG, 1998), *Concurso TVE* (MPEG, 1998), *Inspector Gadget* (imdb, 2005) e *Other Side Of Heaven* (IMDB, 2001). Cada uma destas matrizes é simétrica com dimensão de 6000x6000 pixels. Cada pixel representa o nível de semelhança entre dois pares de imagens do excerto de vídeo. Valores de brilho mais baixo (preto) indicam pares de imagens com elevado grau de similaridade. Valores de brilho mais claros (branco) indicam pares em o grau de similaridade é baixo. De referir que as matrizes são simétricas, ou seja o valor de similaridade da imagem A com a B é idêntico ao da B com a A. Na diagonal os valores de brilho são zero, ou seja indicam que a imagem A é totalmente similar com ela própria.

A.1 Excerto vídeo *Animals*

Descritor *Color Layout*

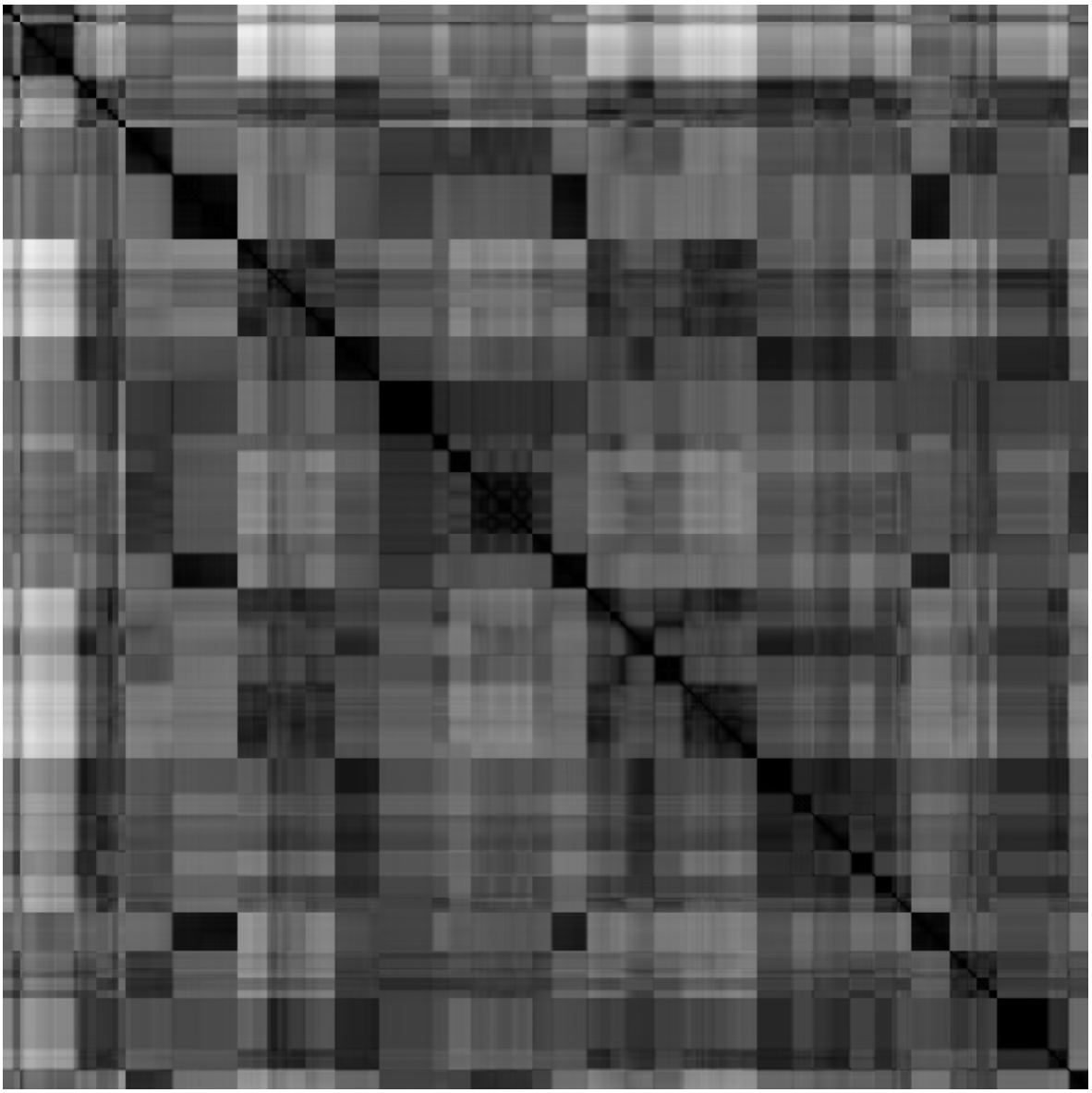


Figura A.1: Matriz de similaridades par a par para o descritor *Color Layout*, excerto vídeo *Animals*

Descritor *Edge Histogram*

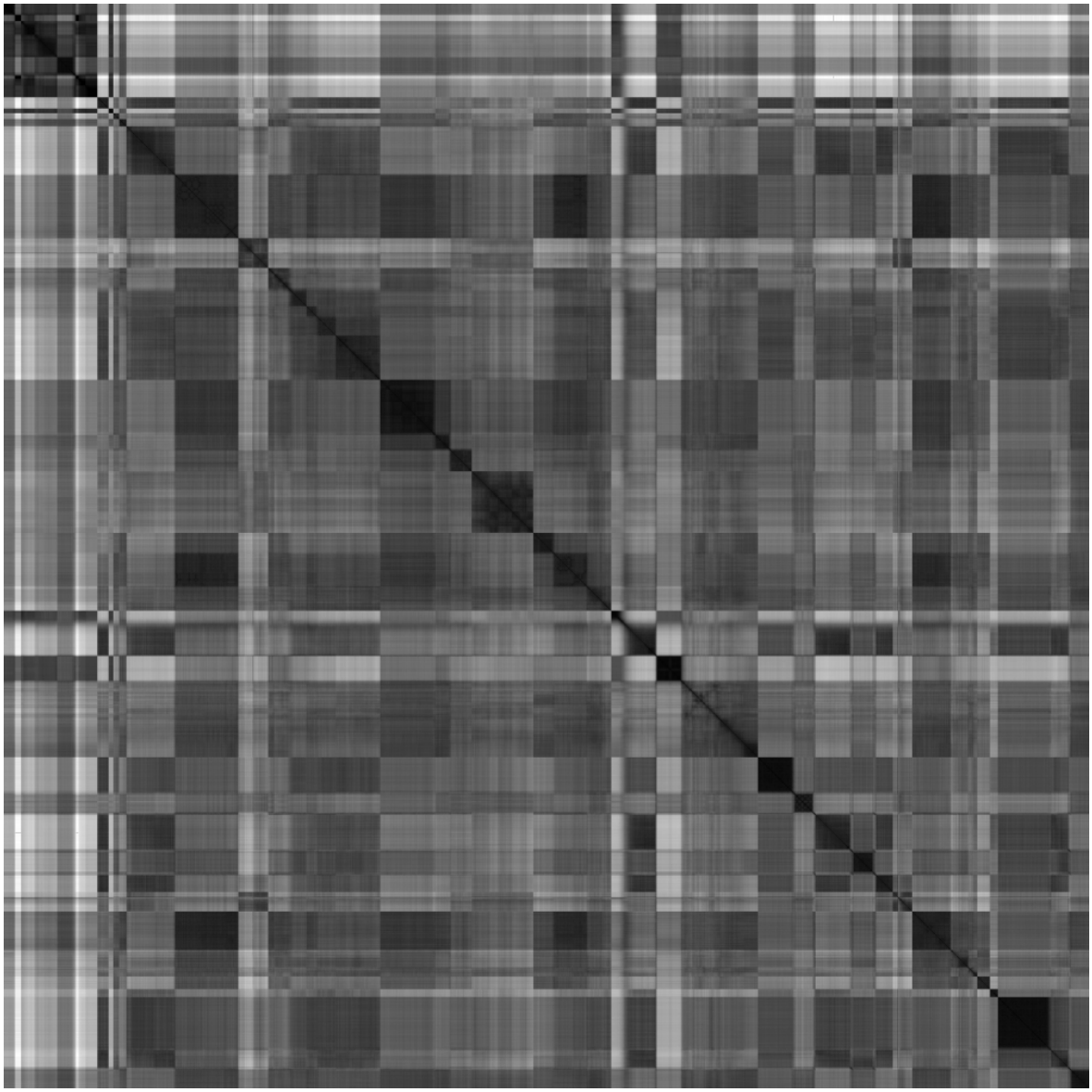


Figura A.2: Matriz de similaridades par a par para o descritor *Edge Histogram*, excerto vídeo *Animals*

Descritor *Homogeneous Texture*

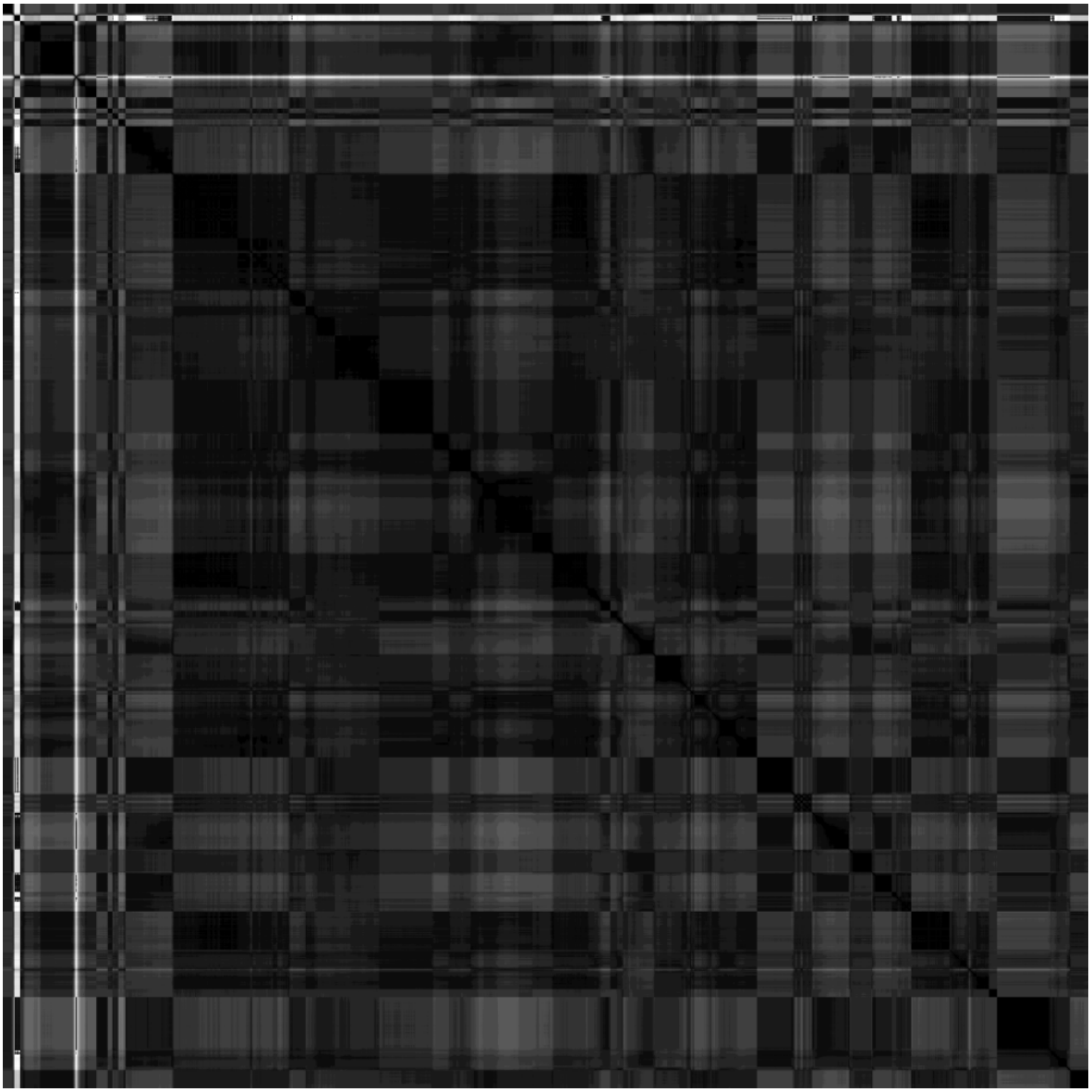


Figura A.3: Matriz de similaridades par a par para o descritor *Homogeneous Texture,* excerto vídeo *Animals*

Descritor *Scalable Color*

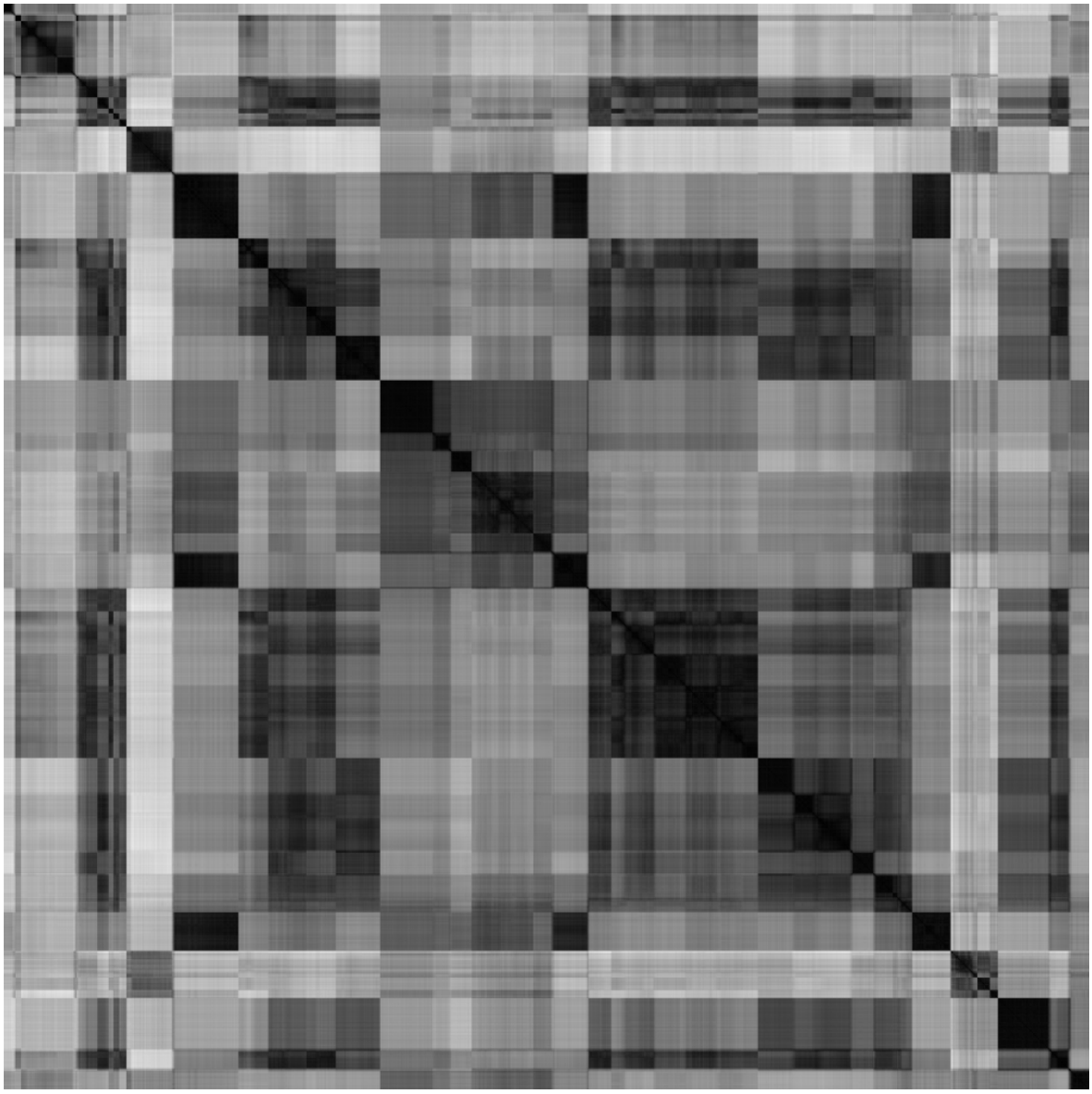


Figura A.4: Matriz de similaridades par a par para o descritor *Scalable Color*, excerto vídeo *Animals*

A.2 Excerto vídeo *Noticias TVE*

Descritor *Color Layout*

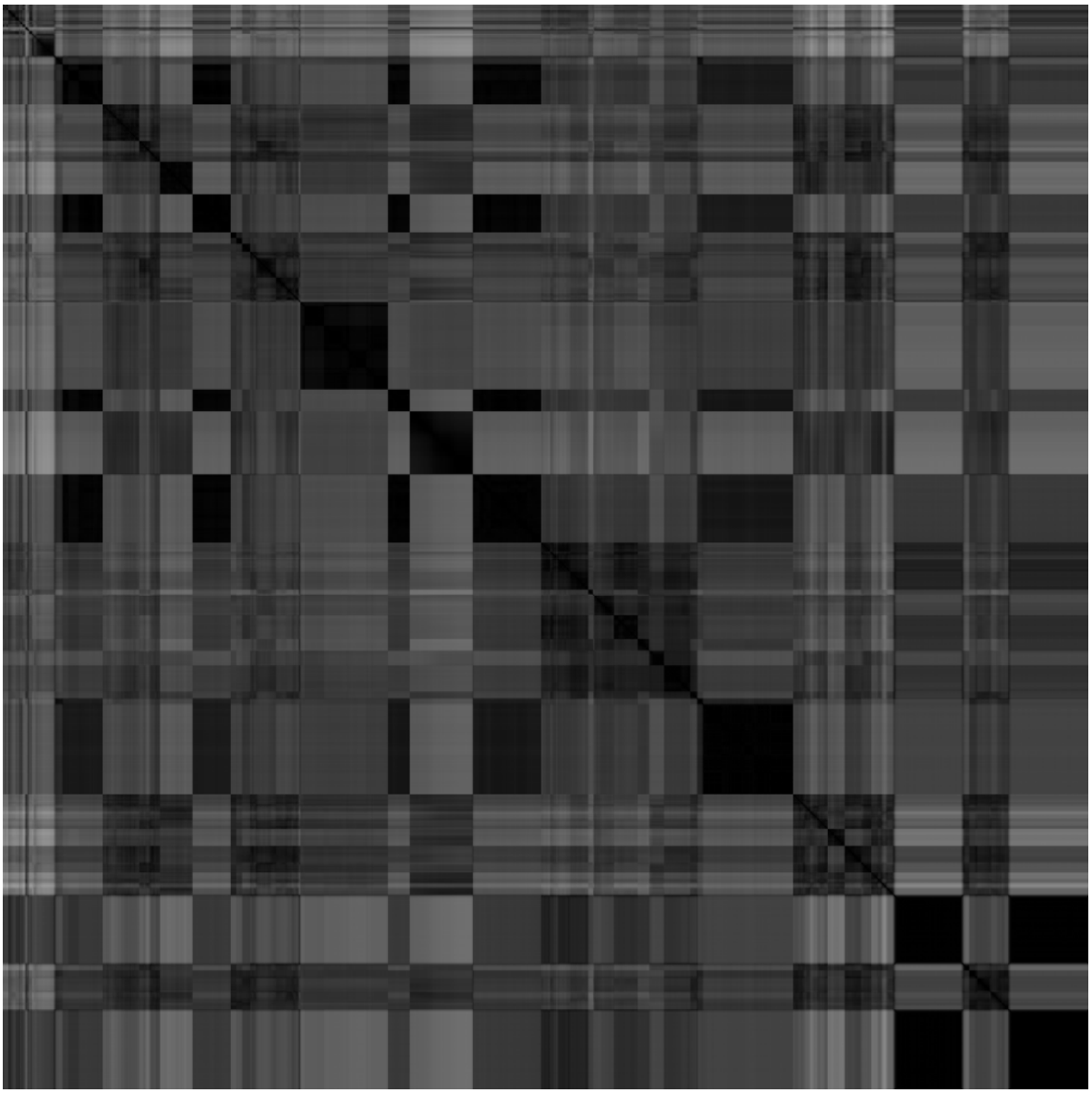


Figura A.5: Matriz de similaridades par a par para o descritor *Color Layout*, excerto vídeo *Noticias TVE*

Descritor *Edge Histogram*

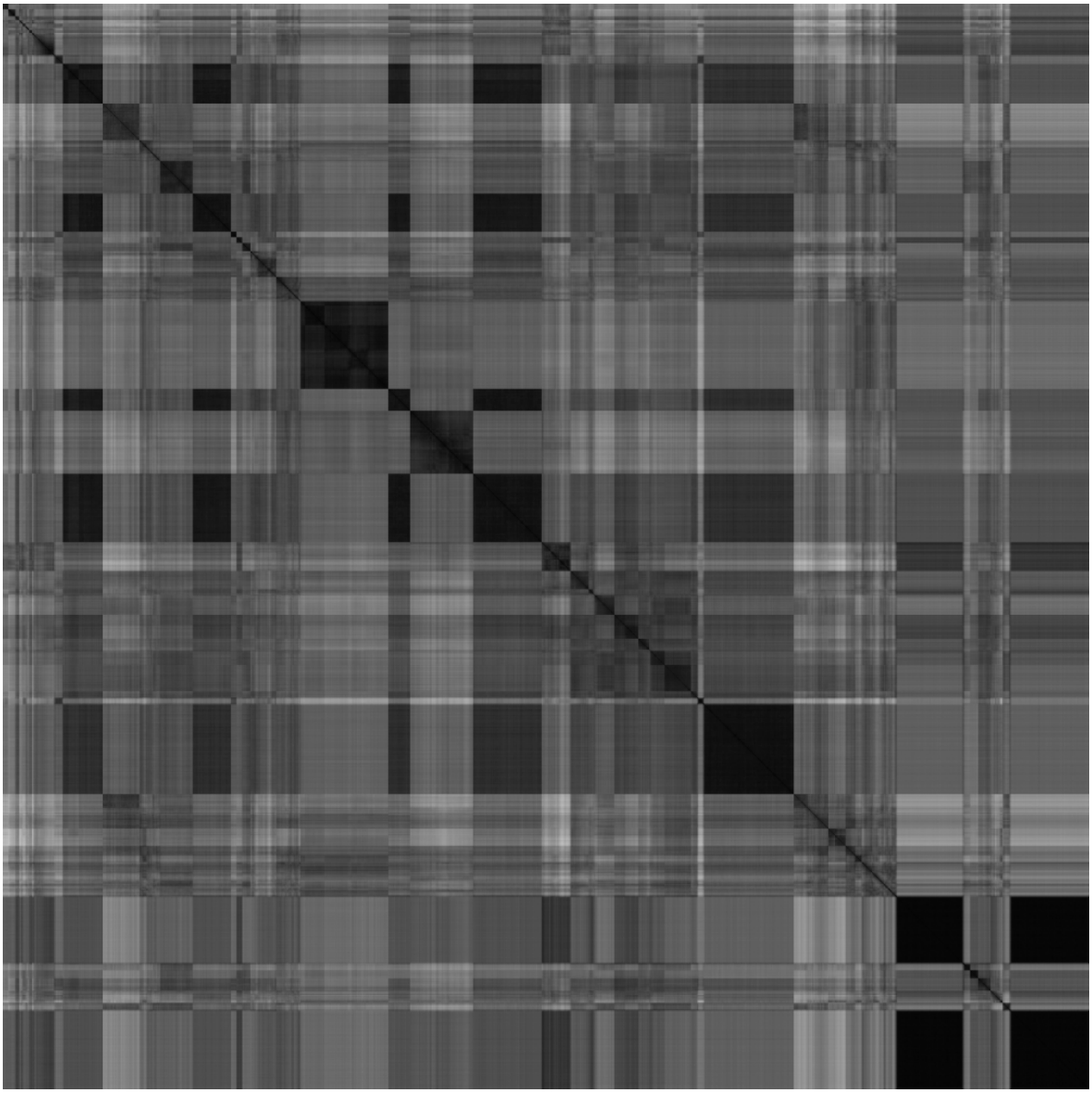


Figura A.6: Matriz de similaridades par a par para o descritor *Edge Histogram*, excerto vídeo *Noticias TVE*

Descritor *Homogeneous Texture*

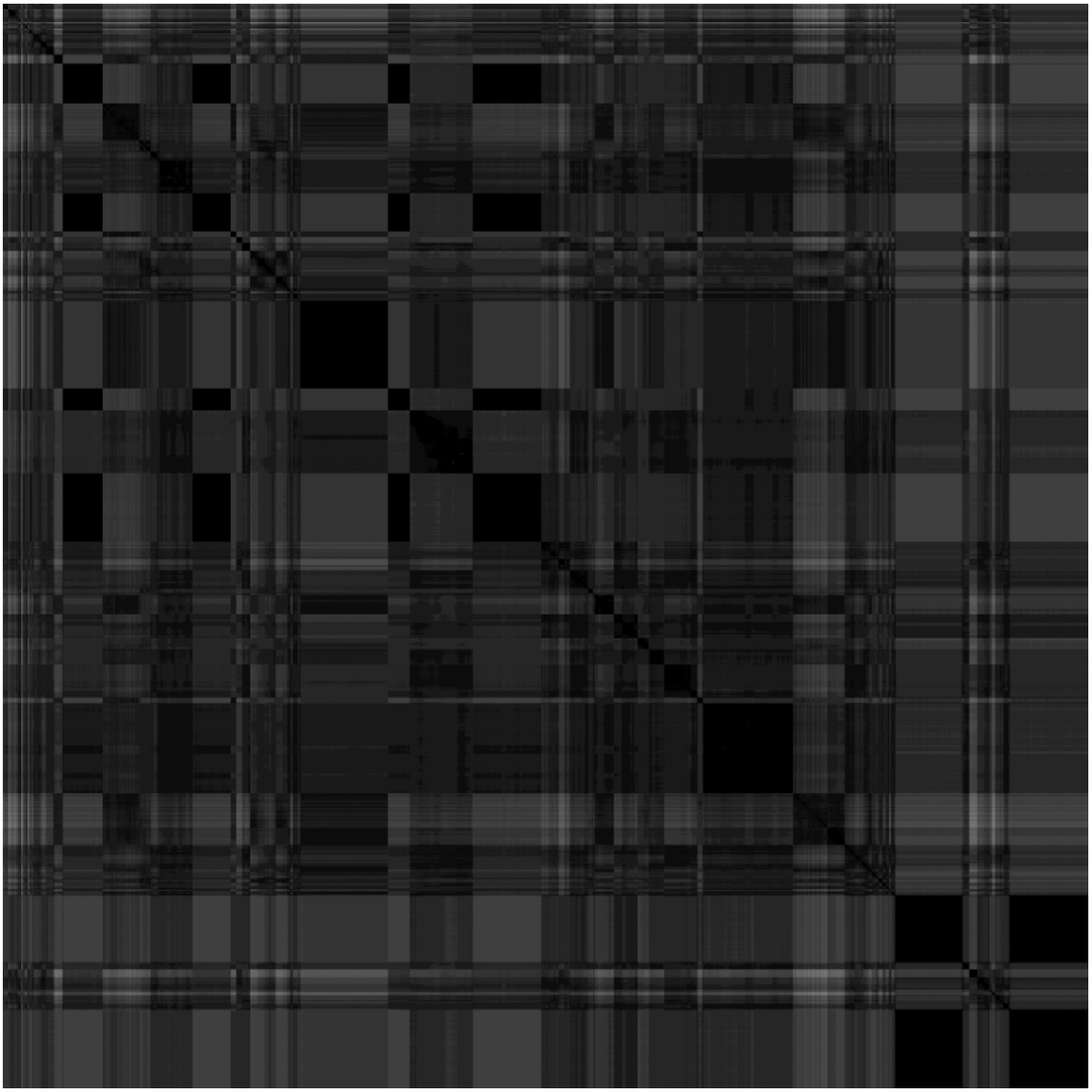


Figura A.7: Matriz de similaridades par a par para o descritor *Homogeneous Texture,* excerto vídeo *Noticias TVE*

Descritor *Scalable Color*

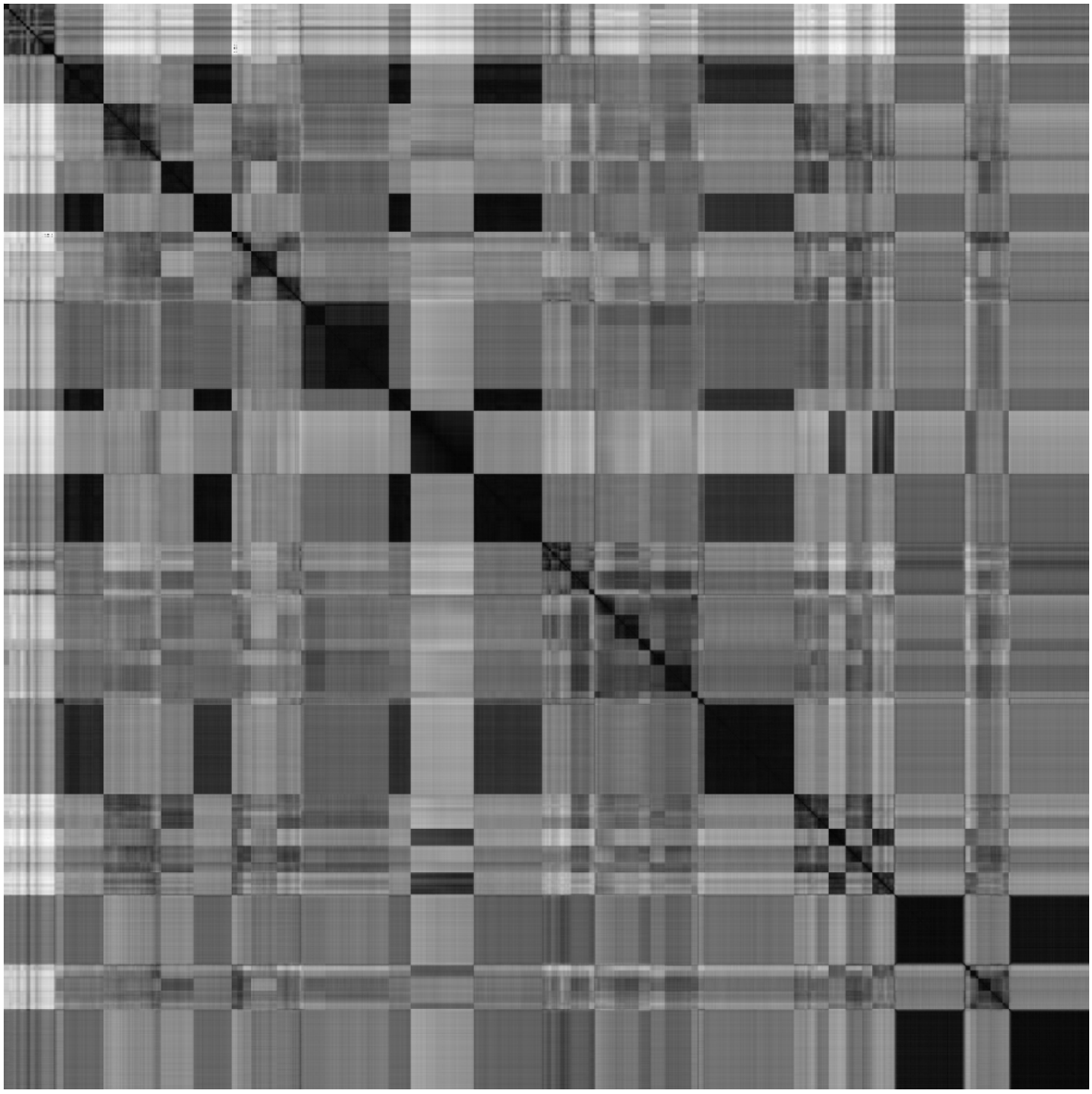


Figura A.8: Matriz de similaridades par a par para o descritor *Scalable Color*, excerto vídeo *Noticias TVE*

A.3 Excerto vídeo *Concurso TVE*

Descritor *Color Layout*

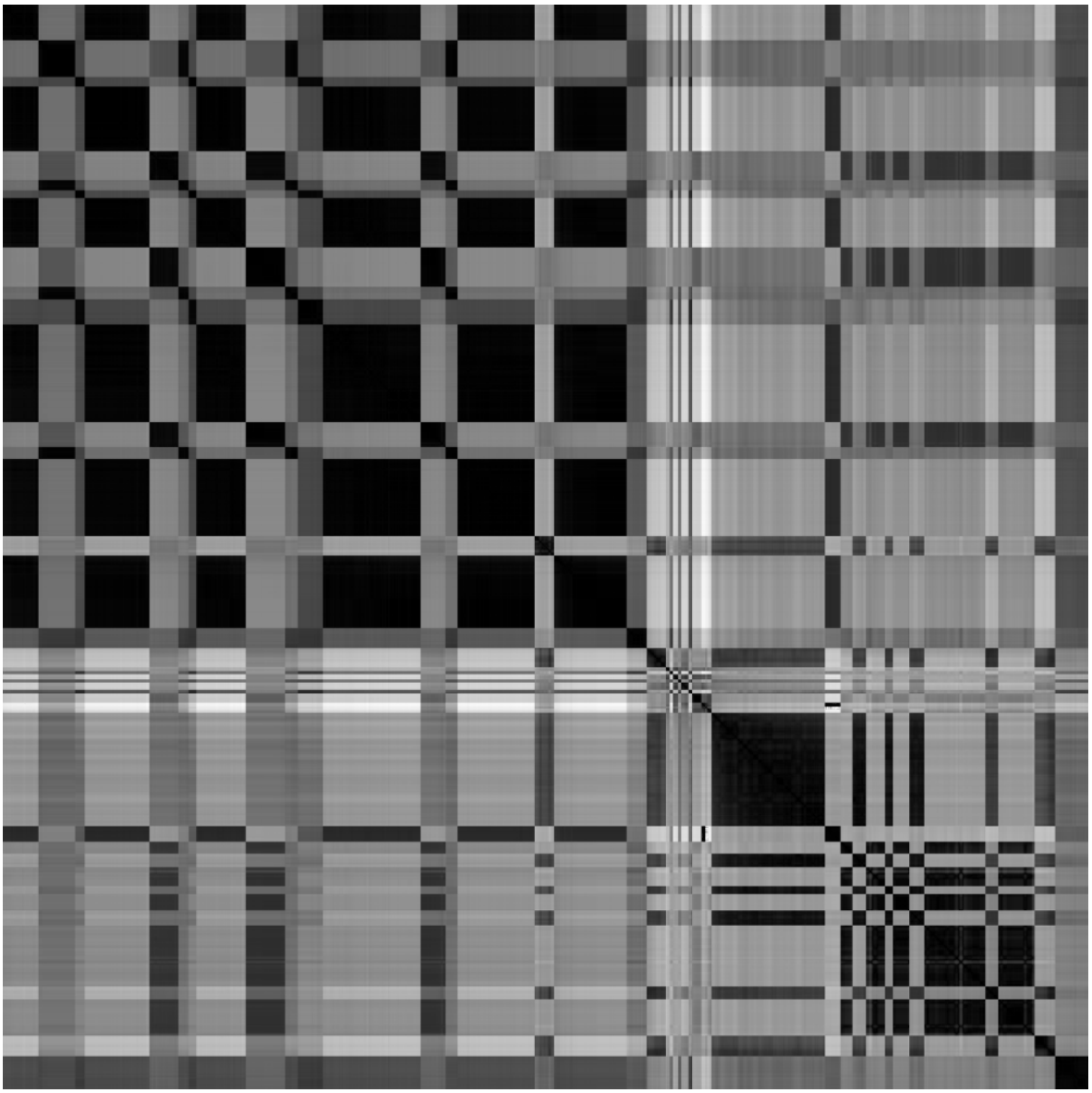


Figura A.9: Matriz de similaridades par a par para o descritor *Color Layout*, excerto vídeo *Concurso TVE*

Descritor *Edge Histogram*

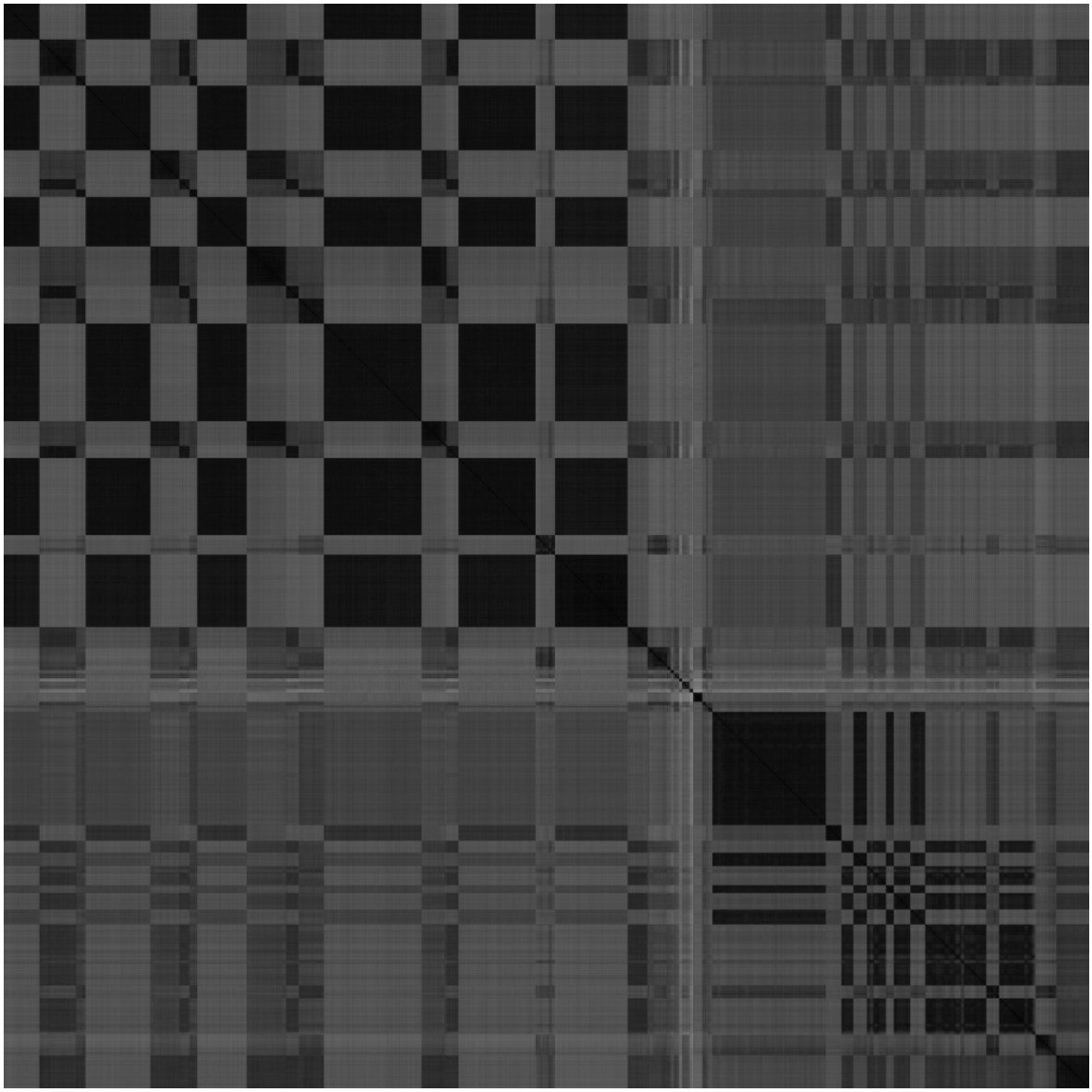


Figura A.10: Matriz de similaridades par a par para o descritor *Edge Histogram*, excerto vídeo *Concurso TVE*

Descritor *Homogeneous Texture*

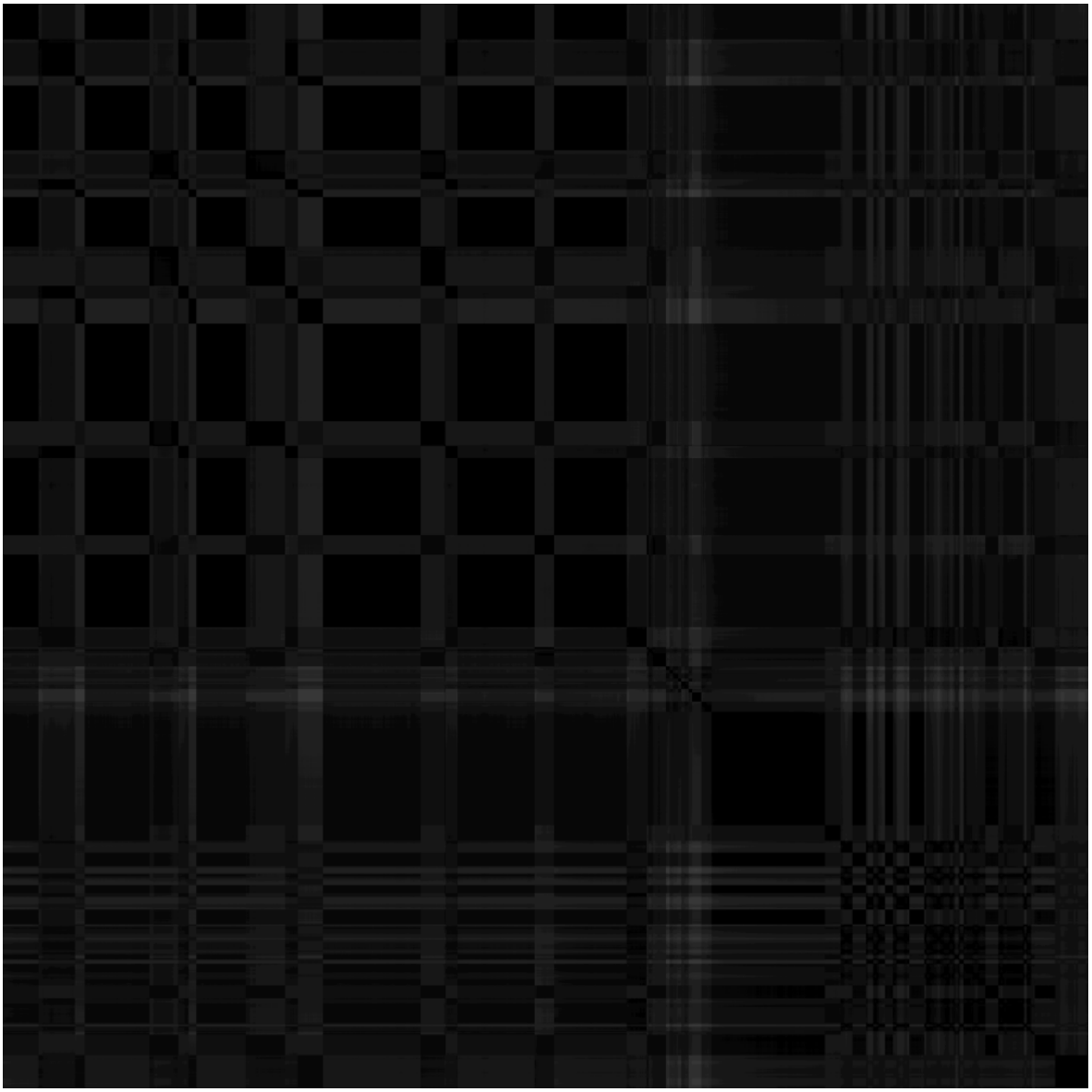


Figura A.11: Matriz de similaridades par a par para o descritor *Homogeneous Texture,* excerto vídeo *Concurso TVE*

Descritor *Scalable Color*

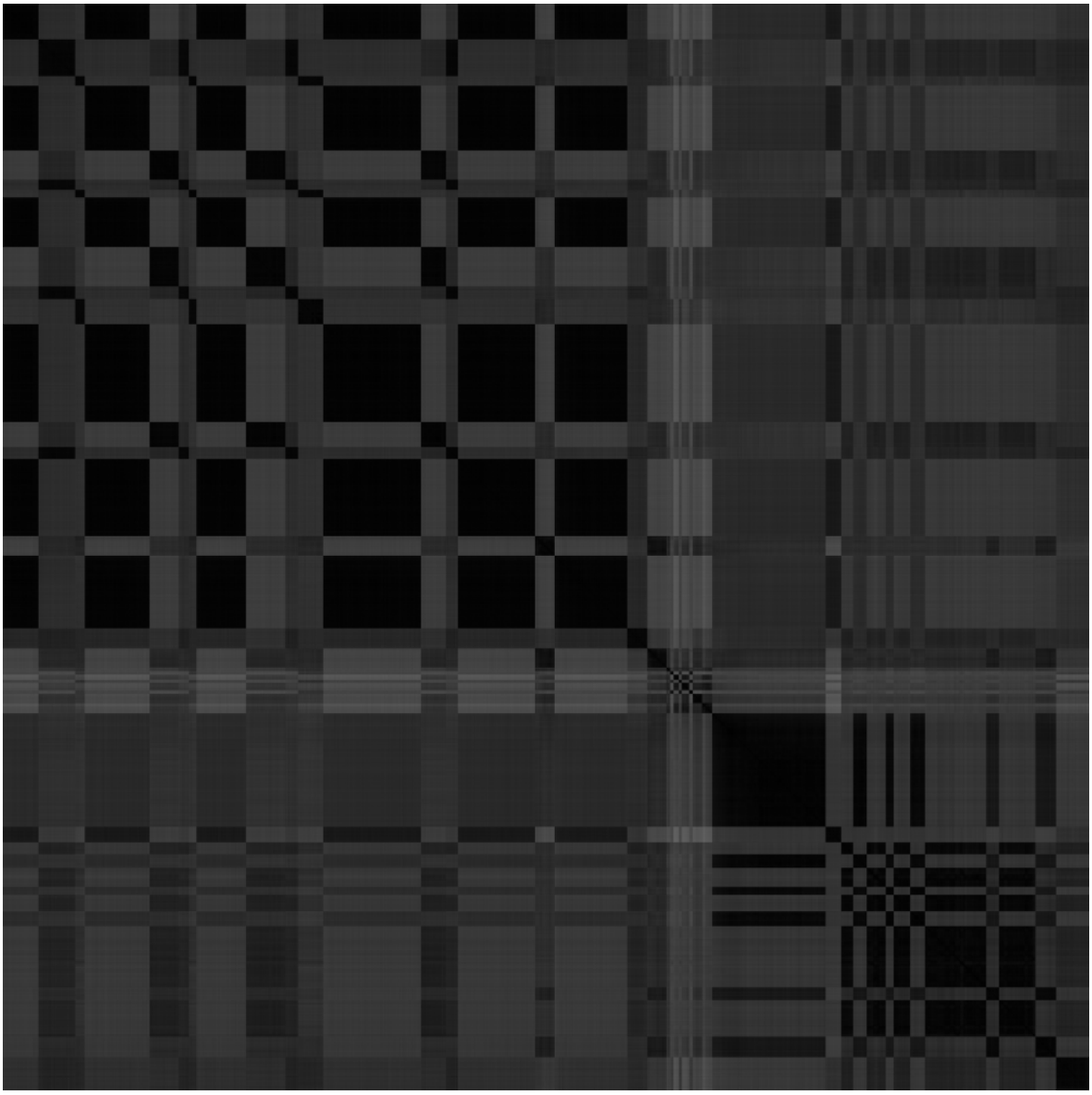


Figura A.12: Matriz de similaridades par a par para o descritor *Scalable Color*, excerto vídeo *Concurso TVE*

A.4 Excerto vídeo *Inspector Gadget*

Descritor *Color Layout*

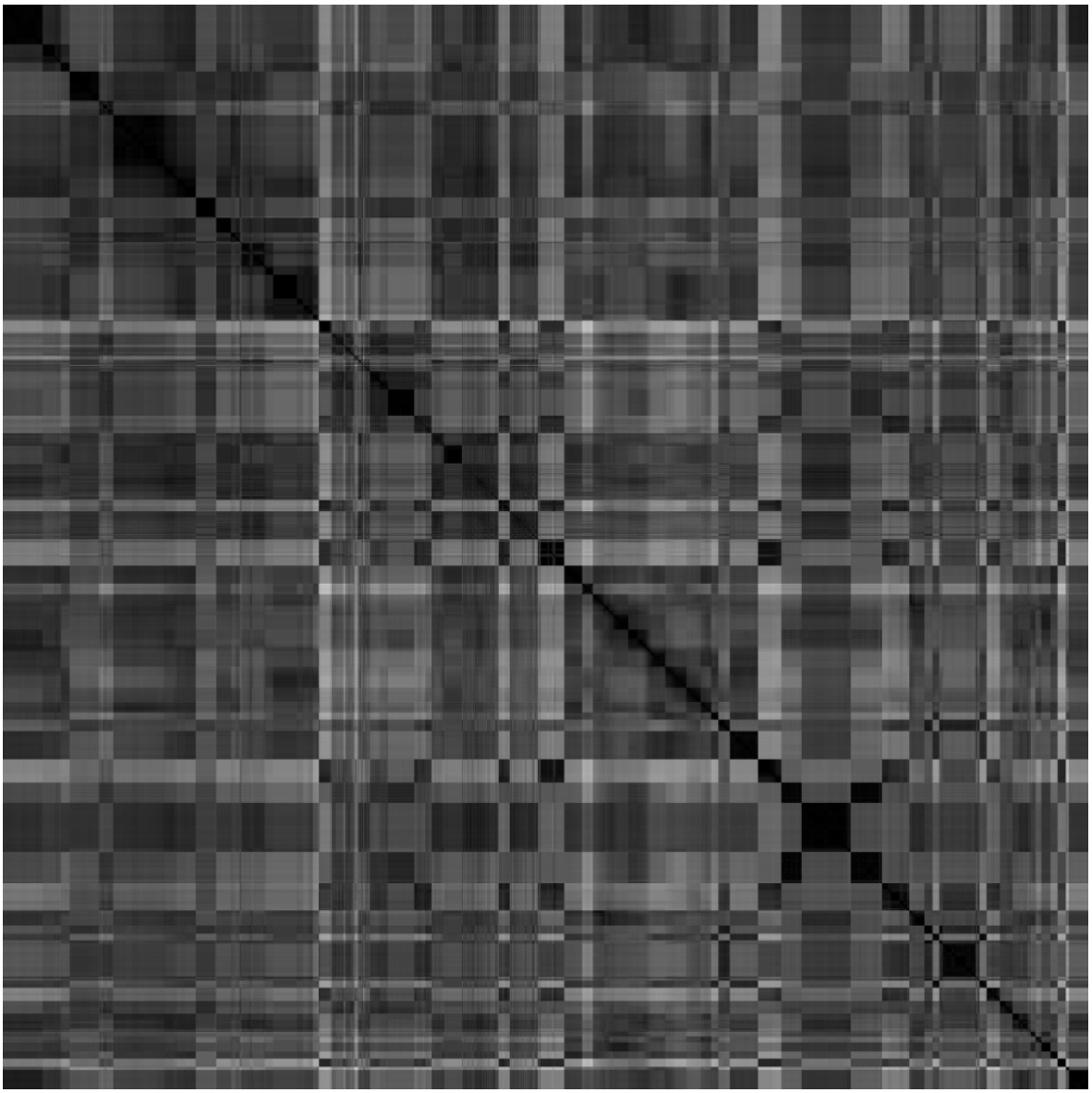


Figura A.13: Matriz de similaridades par a par para o descritor *Color Layout*, excerto vídeo *Inspector Gadget*

Descritor *Edge Histogram*

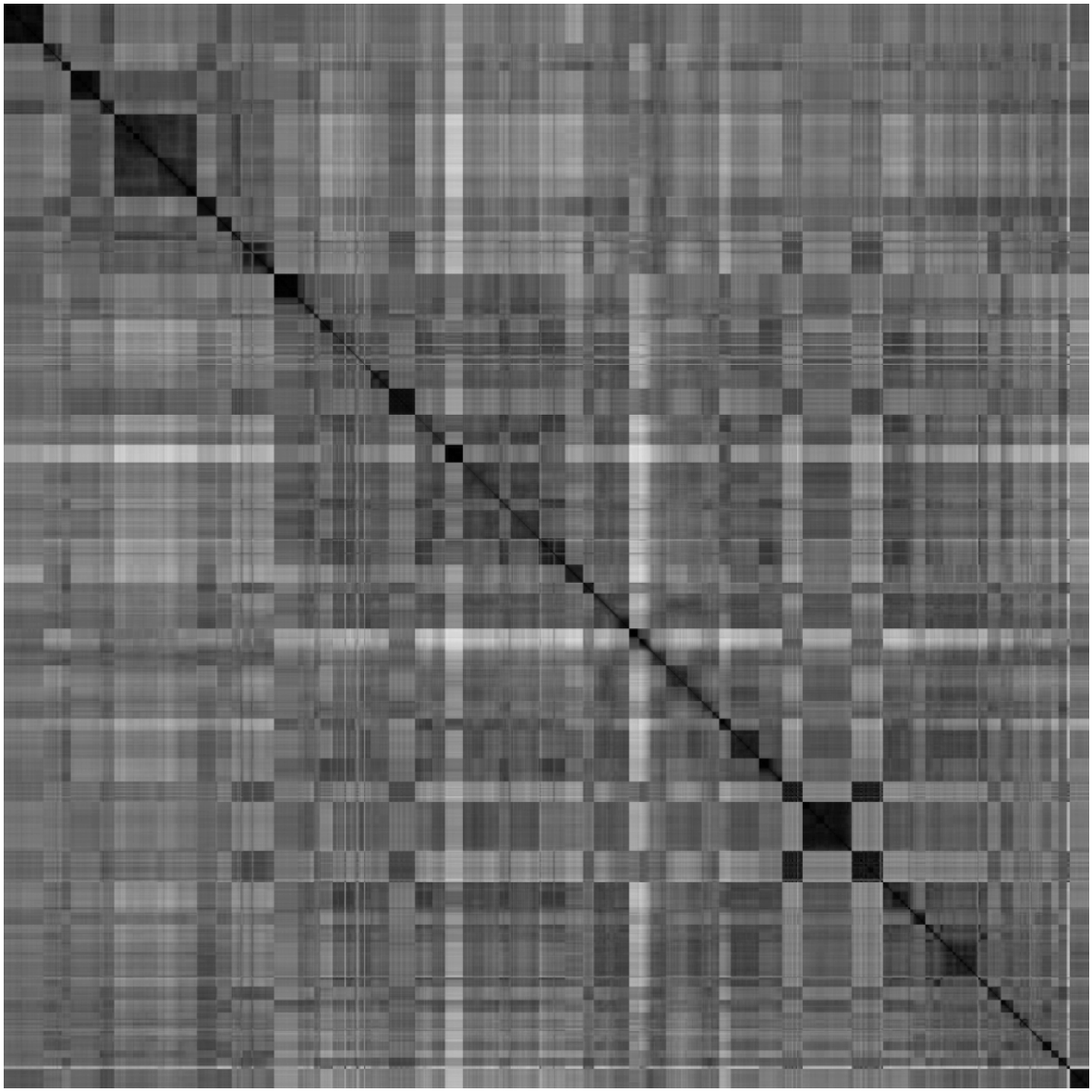


Figura A.14: Matriz de similaridades par a par para o descritor *Edge Histogram*, excerto vídeo *Inspector Gadget*

Descritor *Homogeneous Texture*

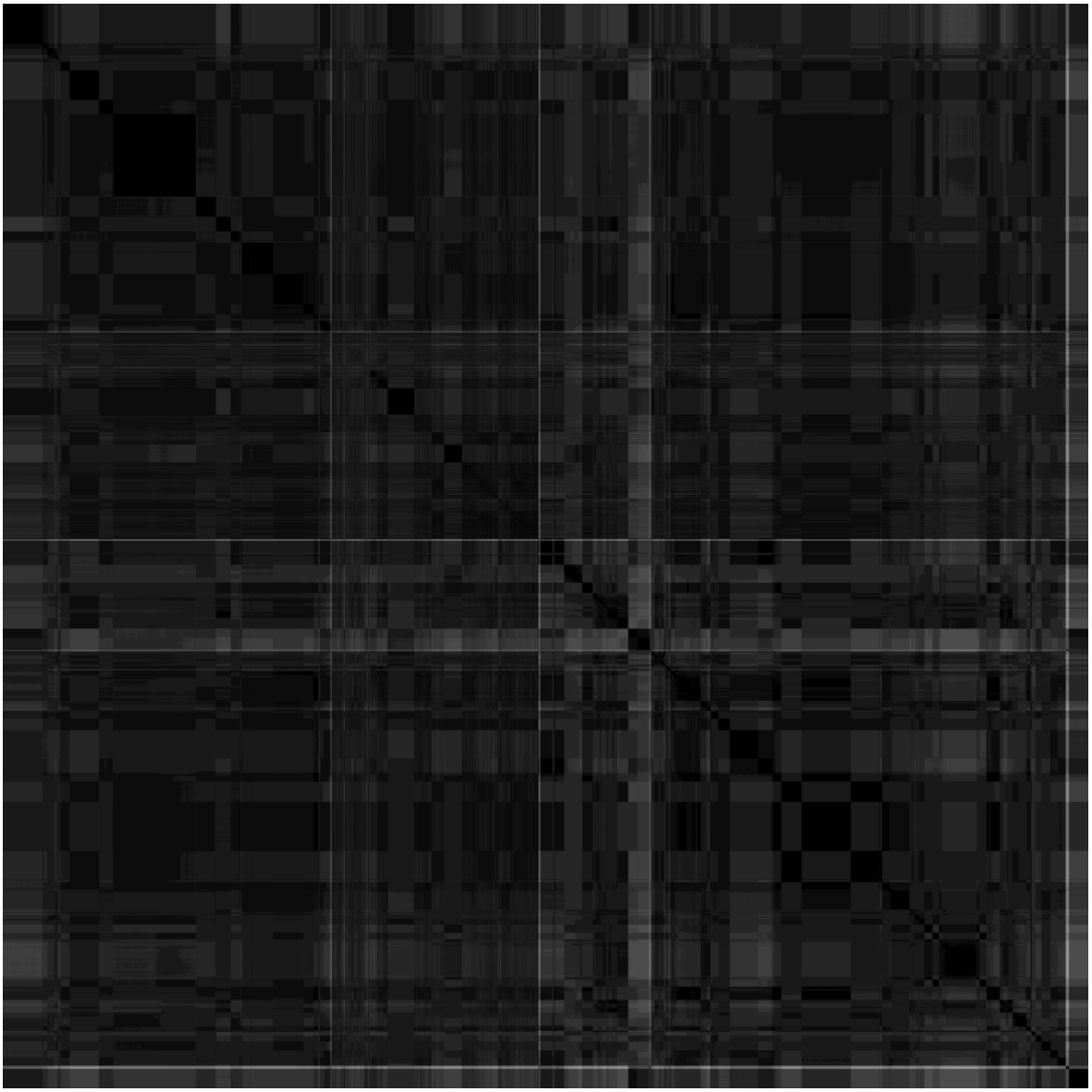


Figura A.15: Matriz de similaridades par a par para o descritor *Homogeneous Texture,* excerto vídeo *Inspector Gadget*

Descritor *Scalable Color*

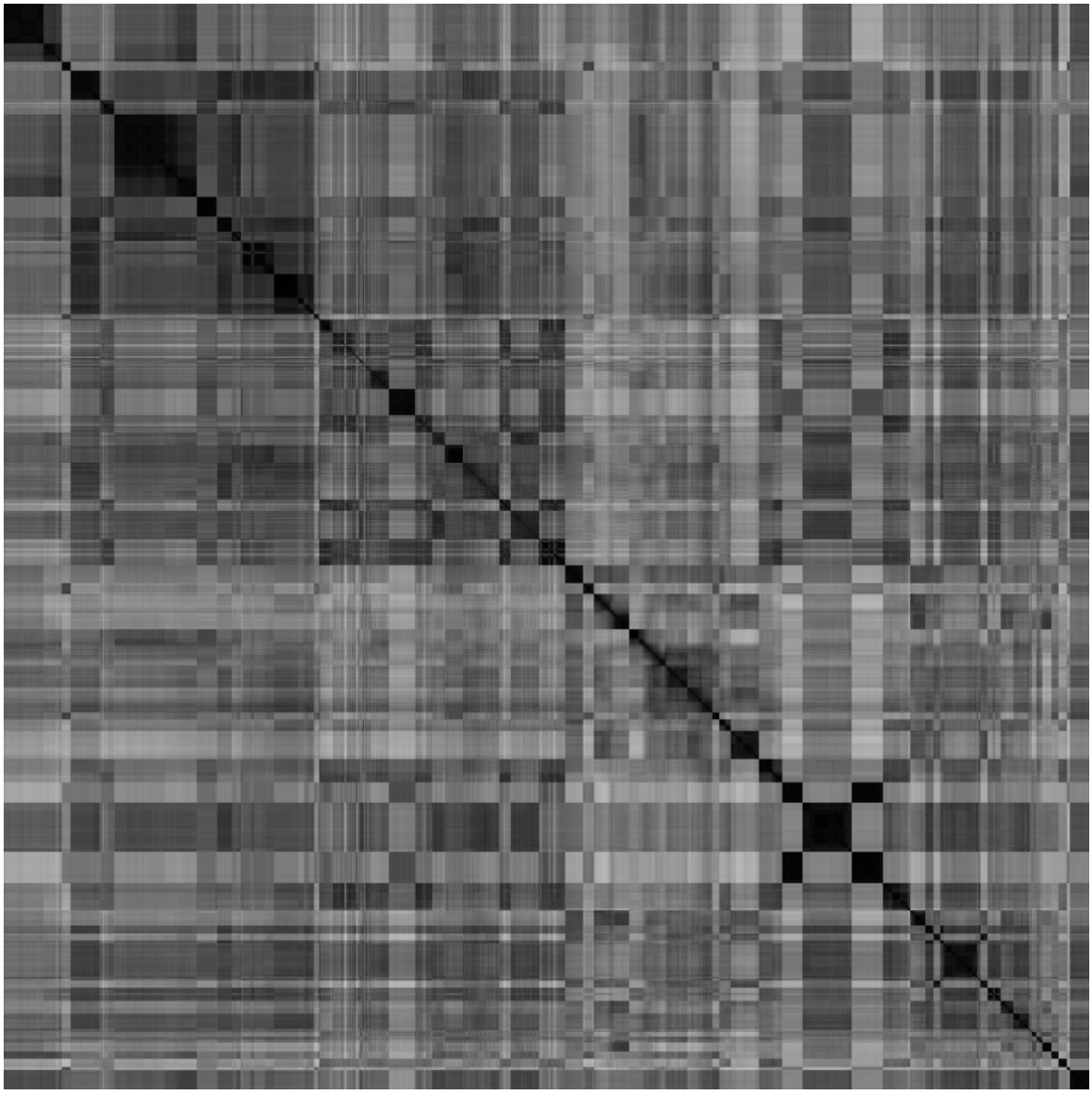


Figura A.16: Matriz de similaridades par a par para o descritor *Scalable Color*, excerto vídeo *Inspector Gadget*

A.5 Excerto vídeo *Other Side Of Heaven*

Descritor *Color Layout*

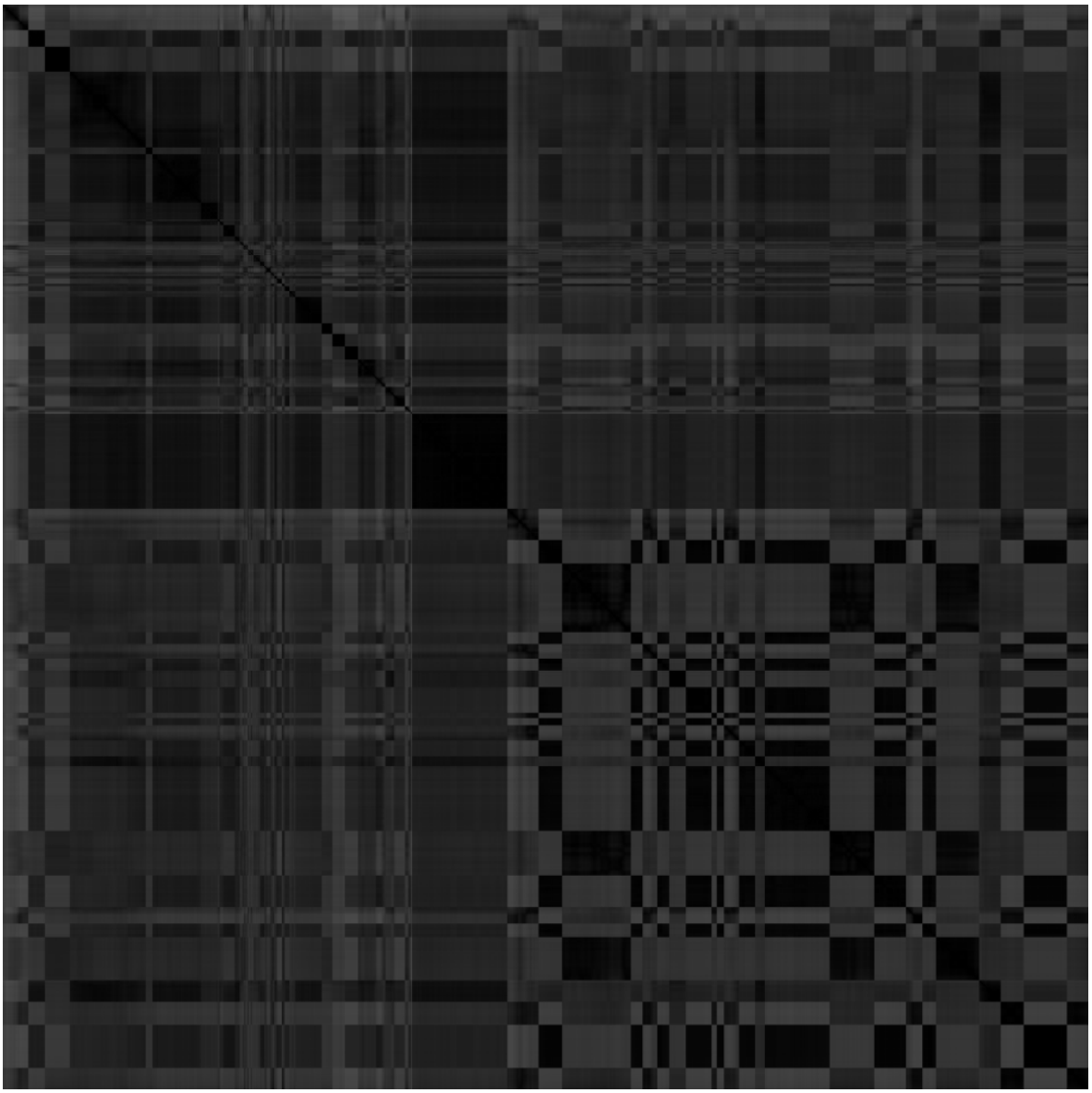


Figura A.17: Matriz de similaridades par a par para o descritor *Color Layout*, excerto vídeo *Other Side Of Heaven*

Descritor *Edge Histogram*

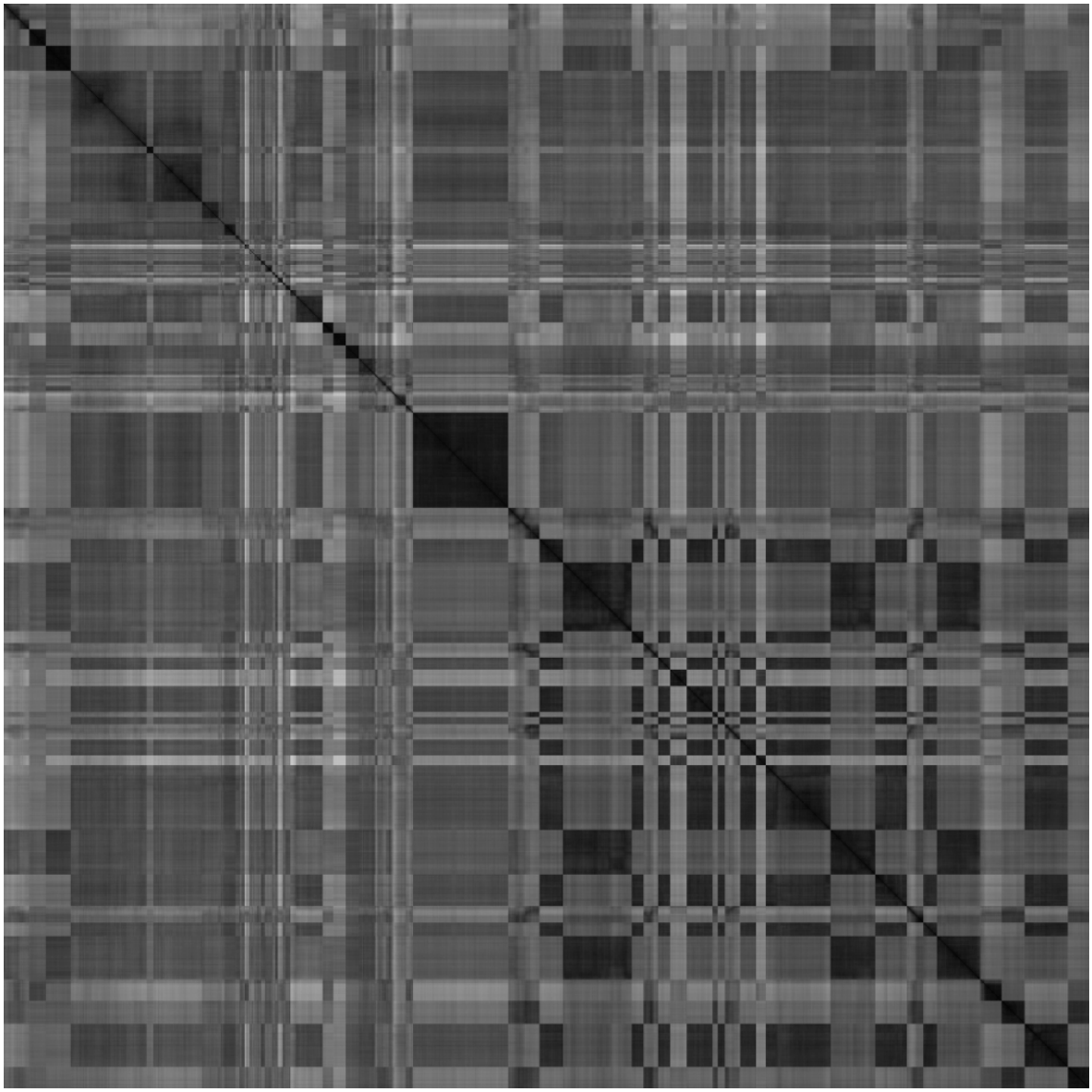


Figura A.18: Matriz de similaridades par a par para o descritor *Edge Histogram*, excerto vídeo *Other Side Of Heaven*

Descritor *Homogeneous Texture*

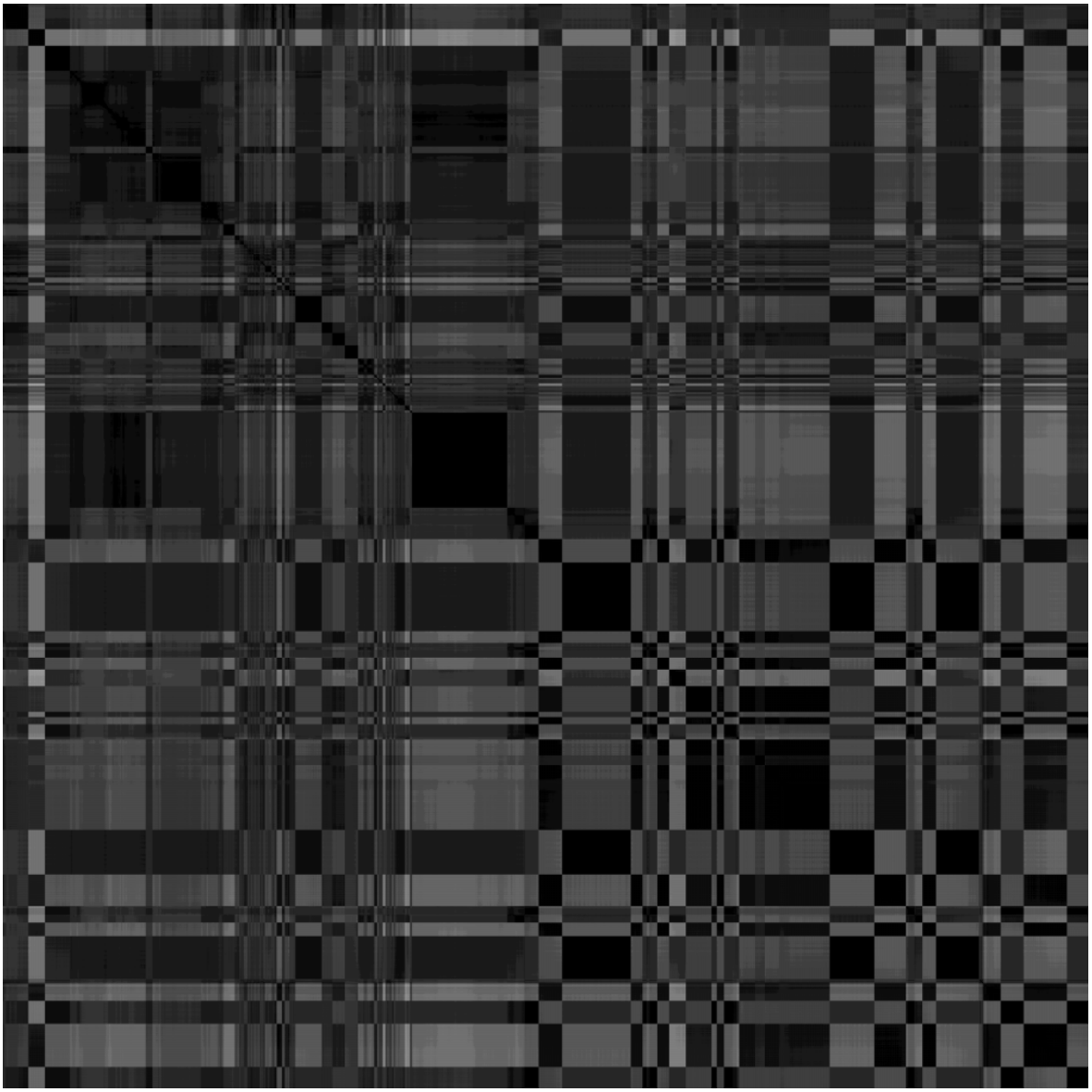


Figura A.19: Matriz de similaridades par a par para o descritor *Homogeneous Texture,* excerto vídeo *Other Side Of Heaven*

Descritor *Scalable Color*

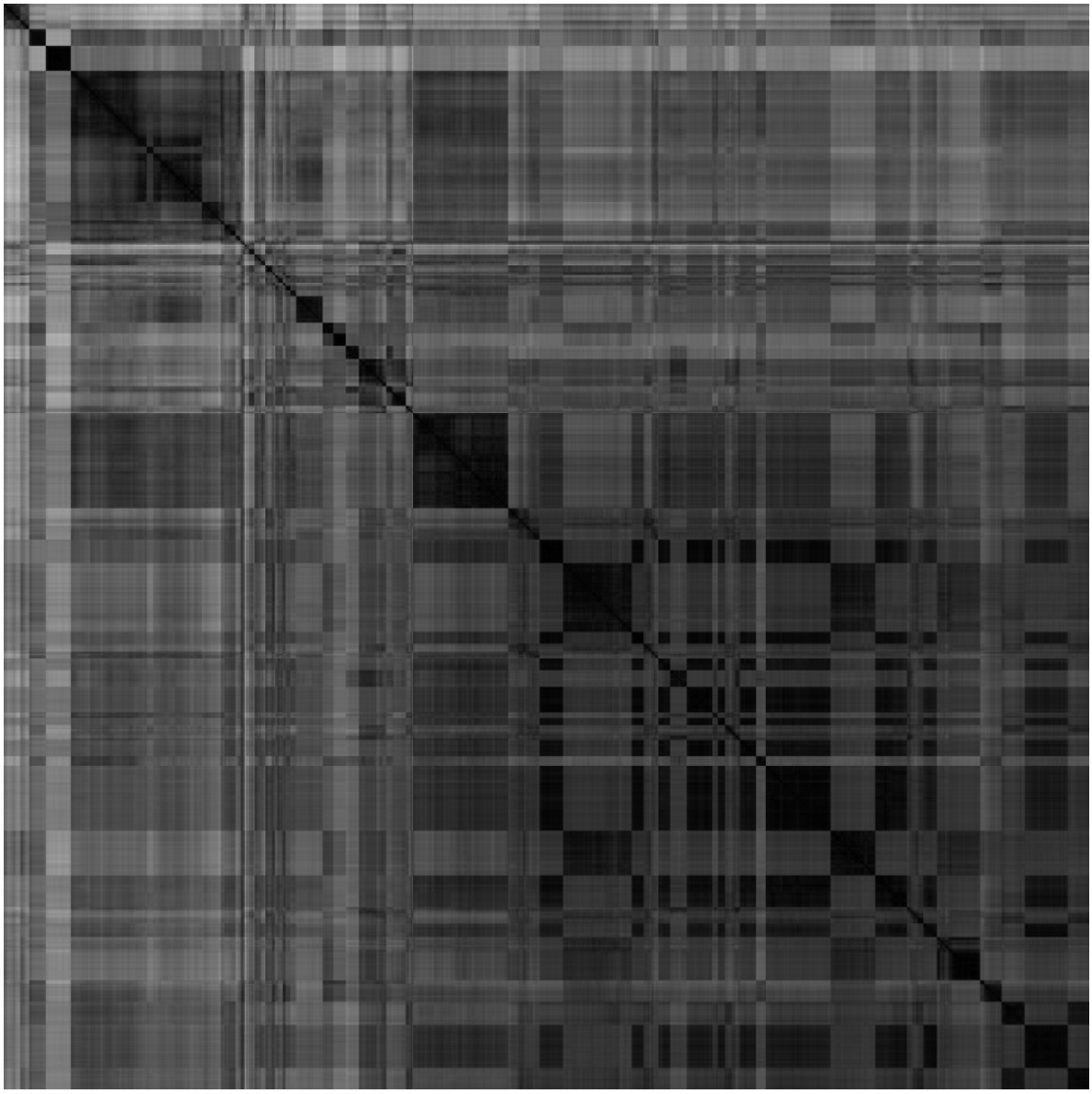


Figura A.20: Matriz de similaridades par a par para o descritor *Scalable Color*, excerto vídeo *Other Side Of Heaven*

Apêndice B

Listagem de Cenas Similares

As secções seguintes ilustram a listagem de todas as cenas contidas em cada um dos excertos de vídeo *Animal* (MPEG, 1998), *Noticias TVE* (MPEG, 1998), *Concurso TVE* (MPEG, 1998), *Inspector Gadget* (imdb, 2005) e *OtherSideOfHeaven* (IMDB, 2001). Para cada uma das cenas são representados na horizontal todos os pares de cenas semelhantes. Esta avaliação de similaridade entre cenas foi feita através de inspecção visual e serve de base à avaliação da identificação de cenas similares no ambiente experimental.

B.1 Excerto vídeo *Animals*

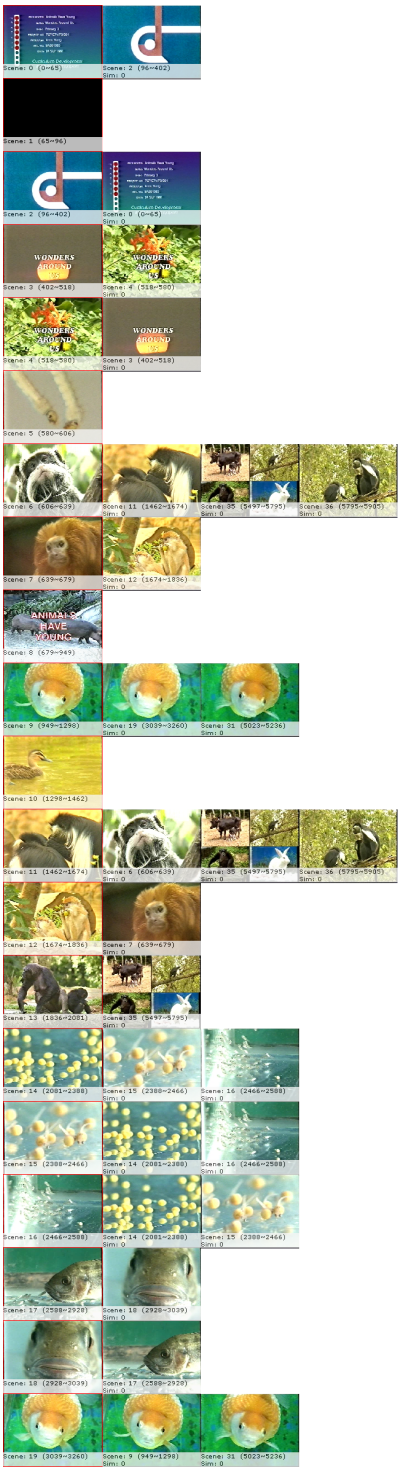


Figura B.1: Lista de cenas similares para o excerto vídeo *Animals* (cenas 1 a 19)

Figura B.2: Lista de cenas similares para o excerto vídeo *Animals* (cenas 20 a 37)

B.2 Excerto vídeo *Noticias TVE*

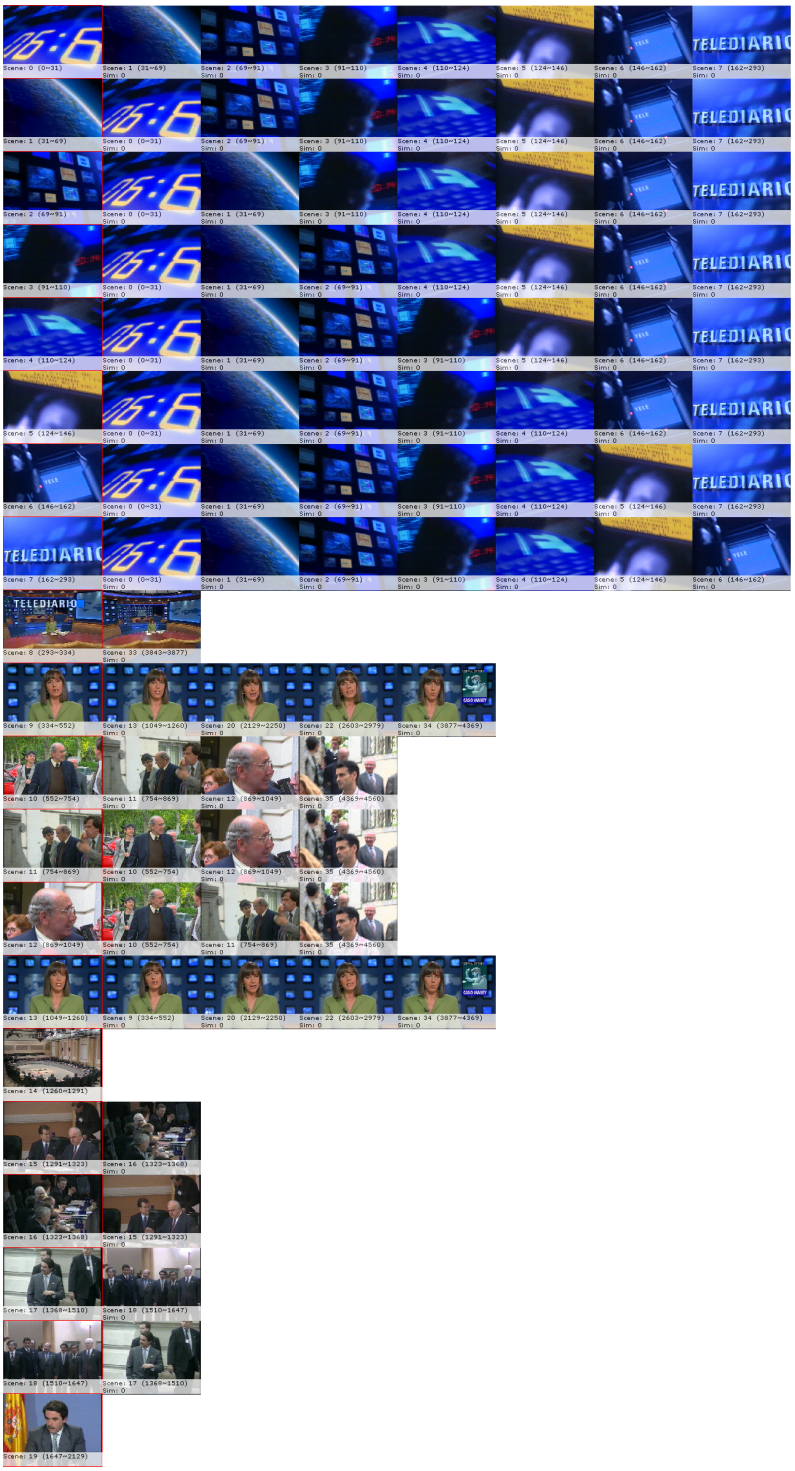


Figura B.3: Lista de cenas similares para o excerto vídeo *Noticias TVE* (cenas 1 a 19)

Figura B.4: Lista de cenas similares para o excerto vídeo *Noticias TVE* (cenas 20 a 39)

Figura B.5: Lista de cenas similares para o excerto vídeo *Noticias TVE* (cenas 40 a 46)

B.3 Excerto vídeo *Concurso TVE*

Figura B.6: Lista de cenas similares para o excerto vídeo *Concurso TVE* (cenas 1 a 19)

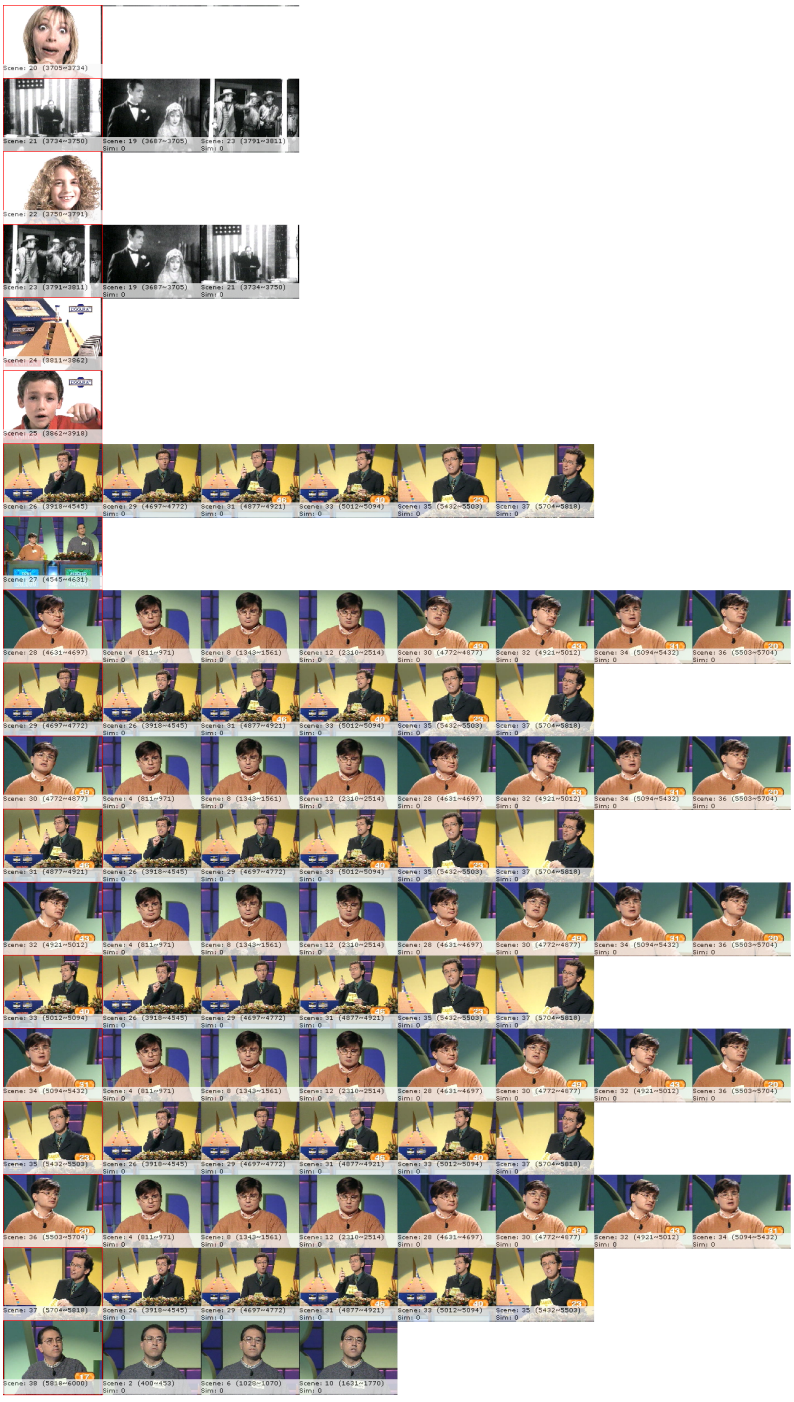


Figura B.7: Lista de cenas similares para o excerto vídeo *Concurso TVE* (cenas 20 a 38)

B.4 Excerto vídeo *Inspector Gadget*

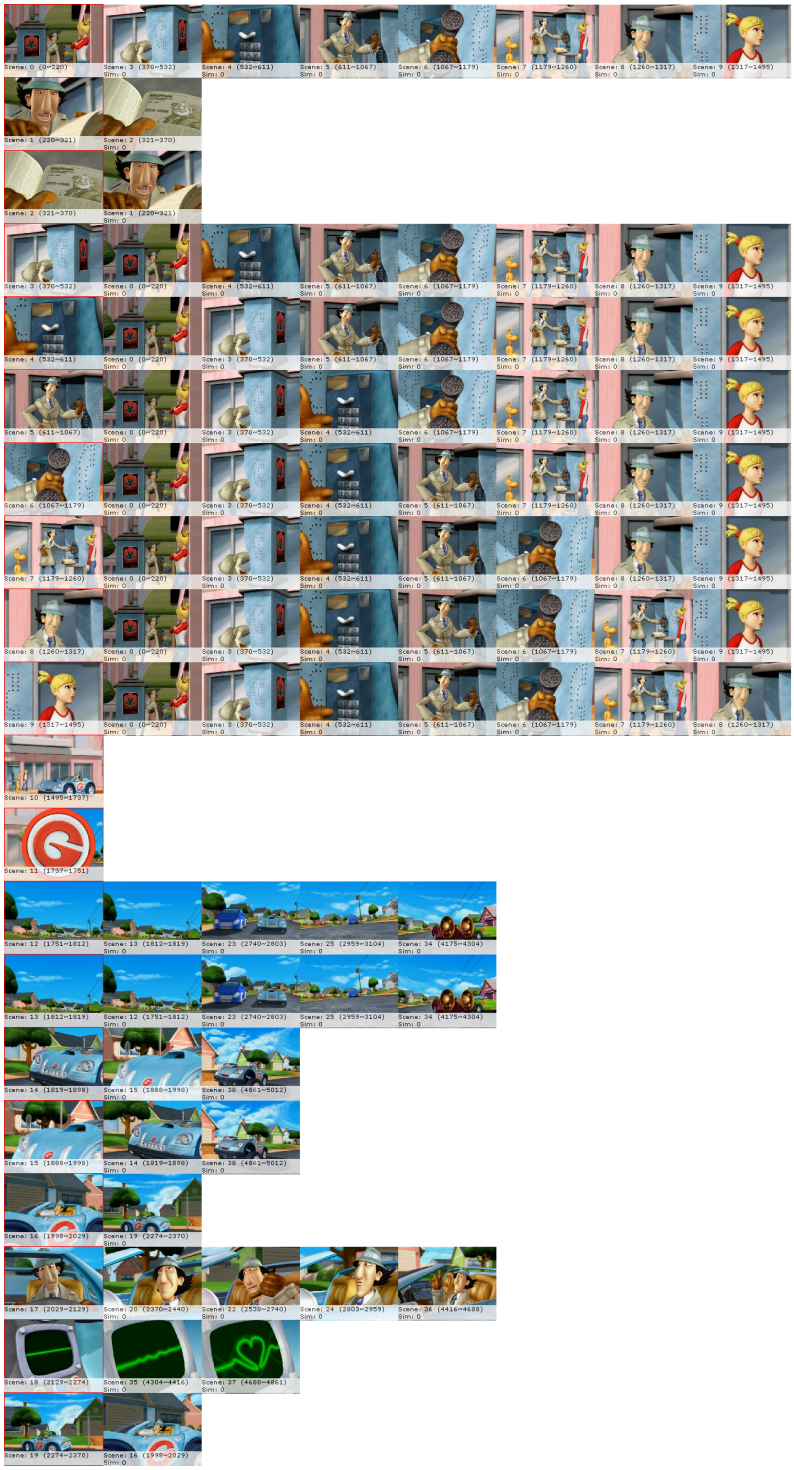


Figura B.8: Lista de cenas similares para o excerto vídeo *Gadget* (cenas 1 a 19)

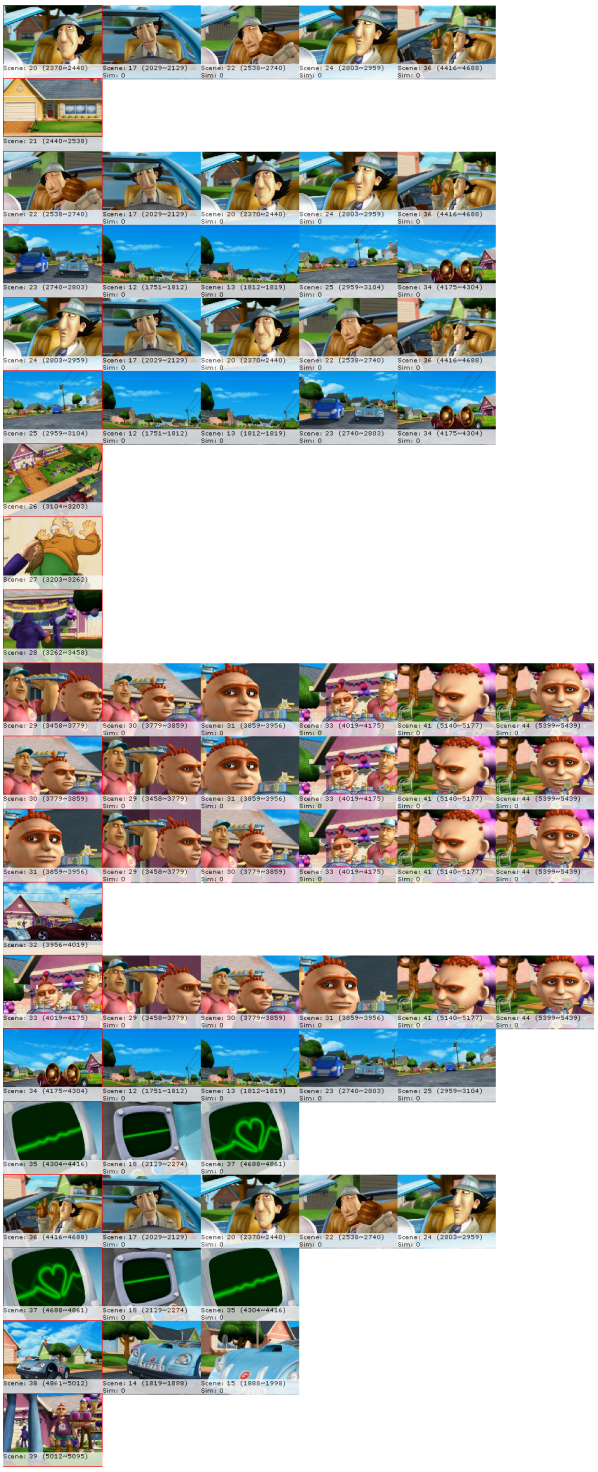


Figura B.9: Lista de cenas similares para o excerto vídeo *Gadget* (cenas 20 a 39)

Figura B.10: Lista de cenas similares para o excerto vídeo *Gadget* (cenas 40 a 55)

B.5 Excerto vídeo *Other Side Of Heaven*

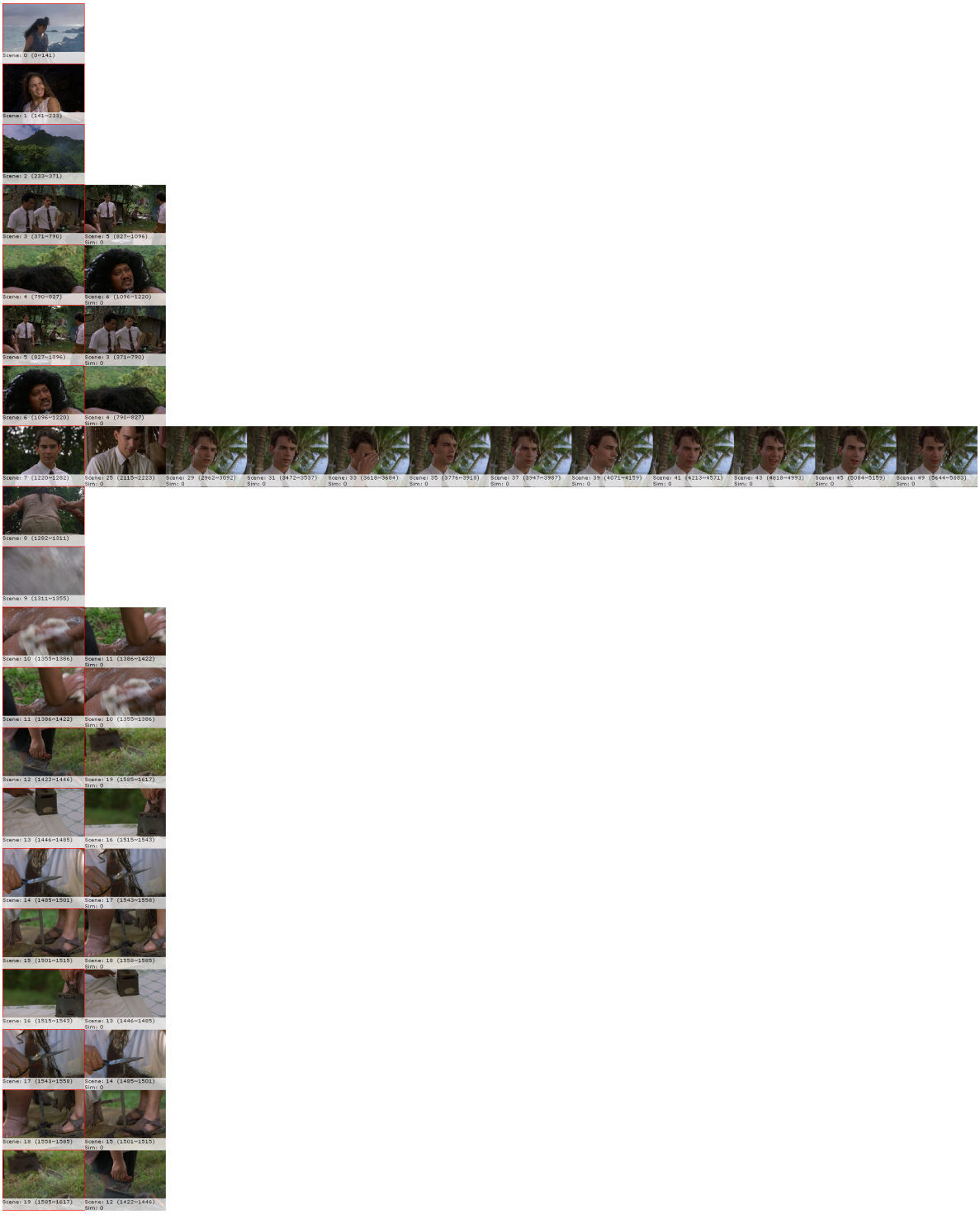


Figura B.11: Lista de cenas similares para o excerto vídeo *Other Side Of Heaven* (cenas 1 a 19)

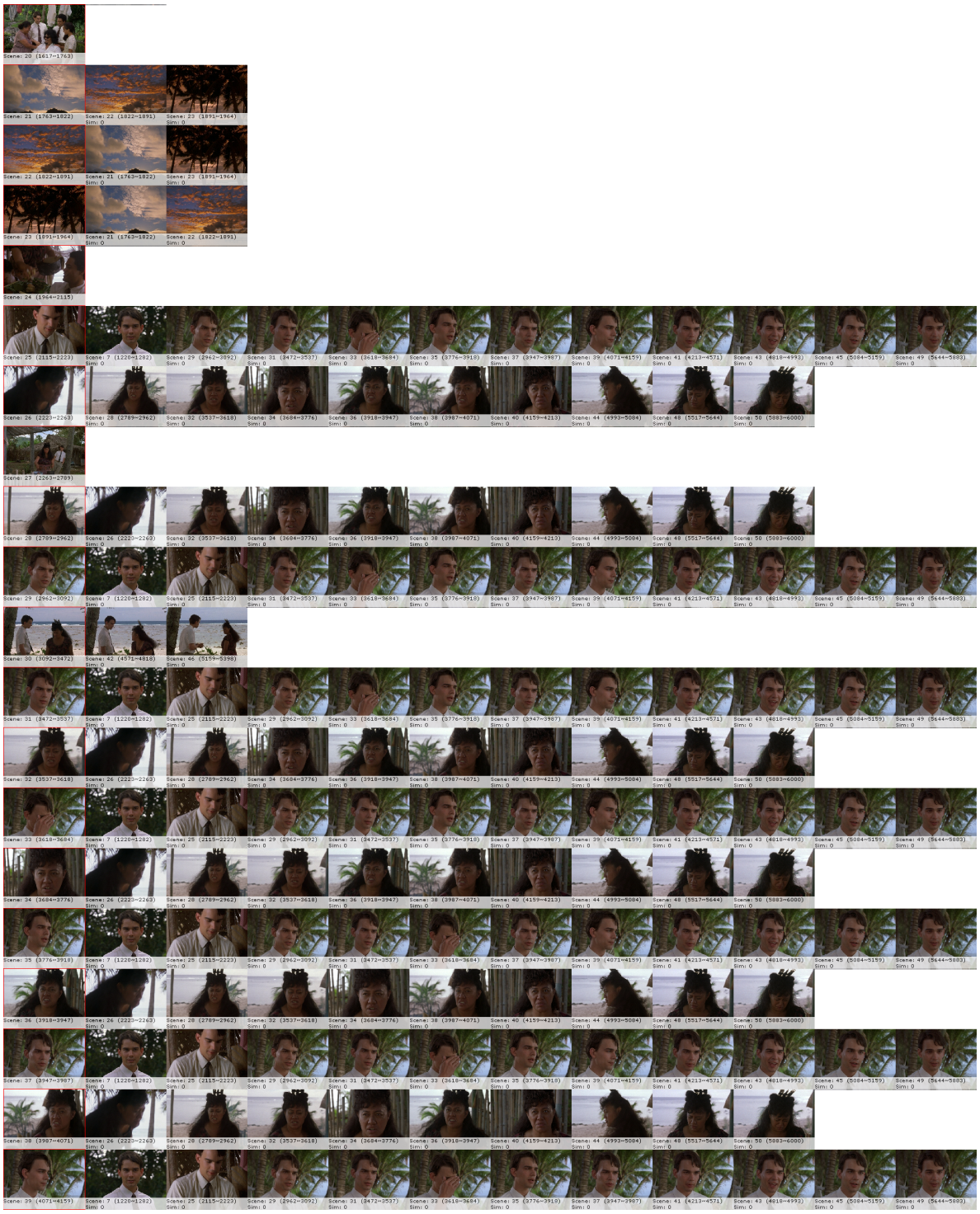


Figura B.12: Lista de cenas similares para o excerto vídeo *Other Side Of Heaven* (cenas 20 a 39)

Figura B.13: Lista de cenas similares para o excerto vídeo *Other Side Of Heaven* (cenas 40 a 50)

Apêndice C

Resultados de Similaridade de Cenas

Neste anexo são incluídos gráficos comparativos dos resultados obtidos na análise de similaridades de cenas. São ilustrados resultados de percentagens de cenas recuperadas agrupadas por intervalos de recuperação e precisão.

Nos gráficos de Comparativos de por tipo de Descritor (C.1) foram considerados 7 intervalos para a recuperação: 0%,]0%,20%],]20%,40%],]40%,60%],]60%,80%],]80%,100%] e 100% a)[1], para a precisão consideraram-se 8 intervalos: 0%,]0%,20%],]20%,40%],]40%,60%],]60%,80%],]80%,100%], 100% b)[2] e 0% c)[3].

Nos gráficos Comparativos por excerto de vídeo e tipo de Descritor (C.2) foram considerados foram considerados 6 intervalos para a recuperação: [0%,20%],]20%,40%],]40%,60%],]60%,80%],]80%,100%] e 100% a), para a precisão consideraram-se 7 intervalos: [0%,20%],]20%,40%],]40%,60%],]60%,80%],]80%,100%], 100% b) e 0% c).

C.1 Comparativos por tipo de Descritor

Neste comparativo podem ser visualizadas as taxas de recuperação e precisão dos descritores *Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Scalable Color* por tipo de

[1]Considerou-se o valor de recuperação a 100% porque a cena não tem pares similares e a resposta tem tamanho Zero

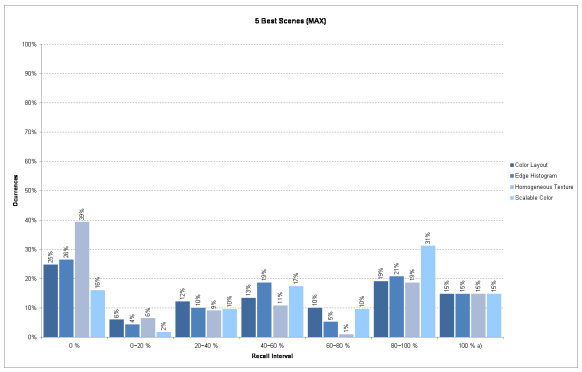
[2]Considerou-se o valor de precisão 100% porque a cena não tem pares similares e a resposta tem tamanho Zero

[3]Considerou-se a precisão 0% porque a cena não tendo pares similares, a resposta tem tamanho diferente de Zero

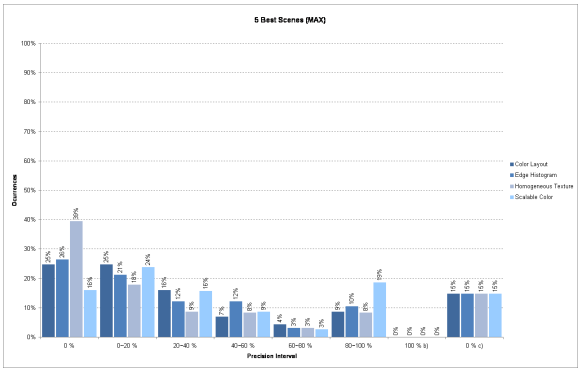
representação de semelhança de cena: Máximo (*MAX*), Mínimo (*MIN*), Média (*MEAN*) e conjunto de recuperação: 5 cenas mais semelhantes (*5 best cenes*), 5 cenas mais semelhantes acima do limiar de semelhança (*5 best cenes above the threshold*) e cenas semelhantes acima do limiar de semelhança (*above the threshold*). Foram agrupados todos os resultados para os excertos de vídeo *Animal* (MPEG, 1998), *Noticias TVE* (MPEG, 1998), *Concurso TVE* (MPEG, 1998), *Inspector Gadget* (imdb, 2005) e *Other Side Of Heaven* (IMDB, 2001) por forma a tirar conclusões gerais do comportamento dos Descritores independentemente do conteúdo.

C.1.1 Análise utilizando o Máximo de semelhança da Cena

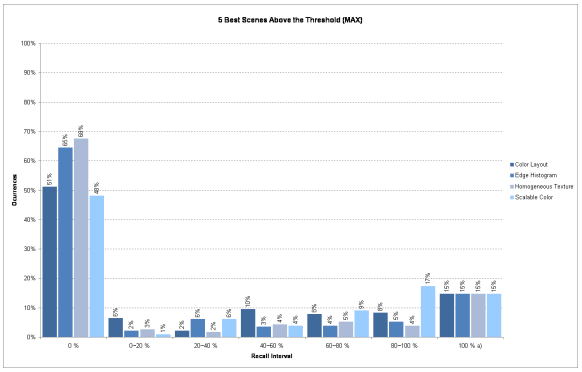
Recuperação | Precisão



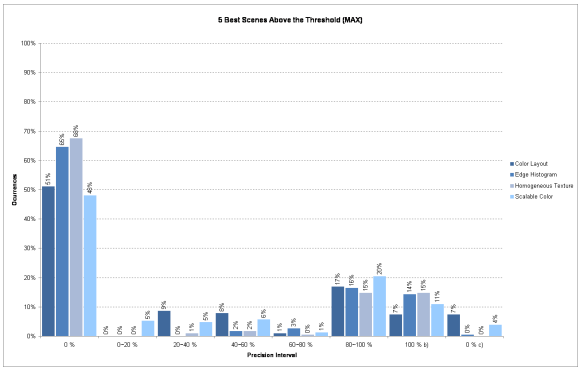
a) 5 Cenas Mais Semelhantes



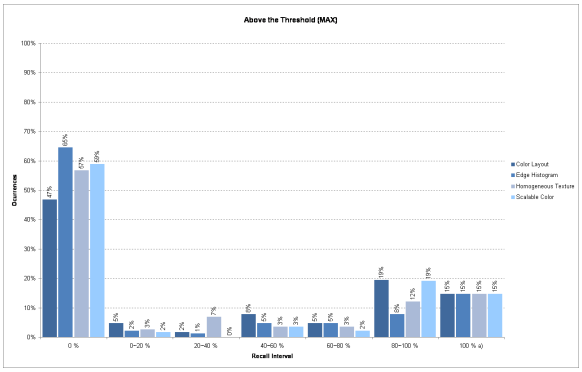
b) 5 Cenas Mais Semelhantes



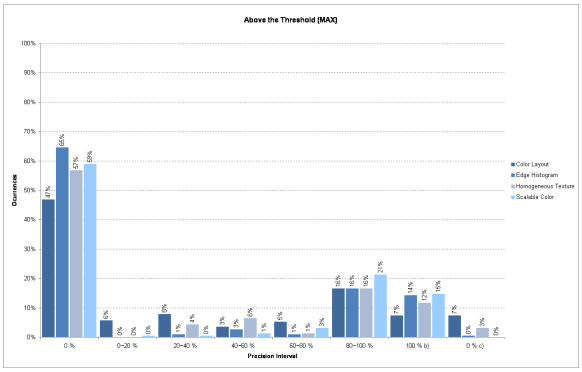
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

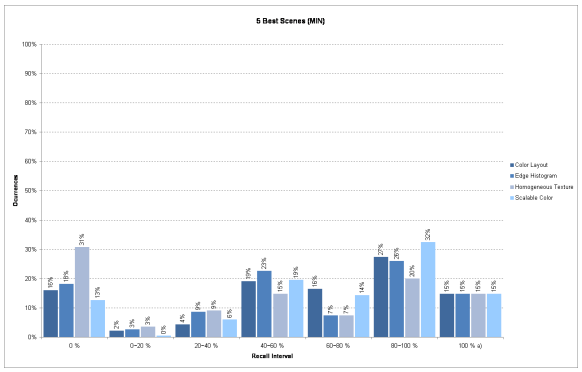


f) Cenas Semelhantes Acima do Limiar de Semelhança

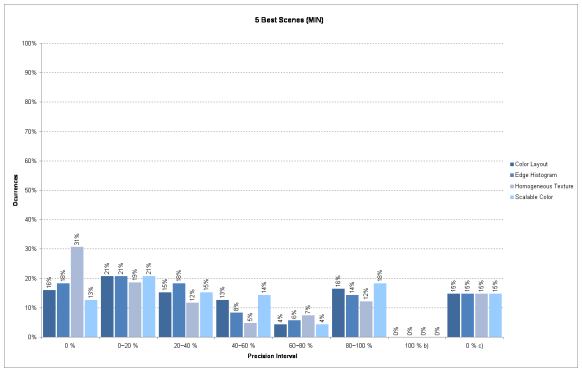
Figura C.1: Taxas de recuperação e precisão utilizando o máximo de semelhança da cena

C.1.2 Análise utilizando o Mínimo de semelhança da Cena

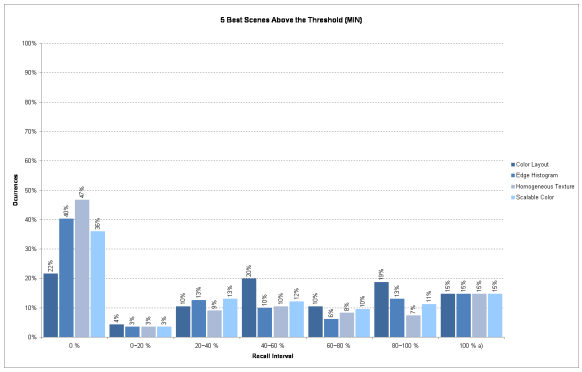
Recuperação | Precisão



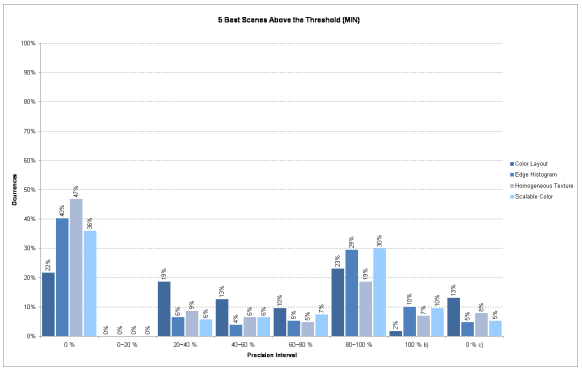
a) 5 Cenas Mais Semelhantes



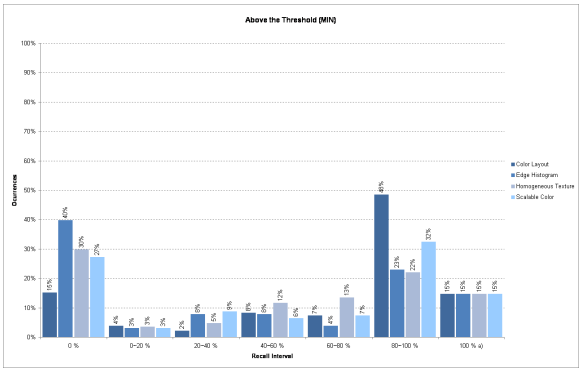
b) 5 Cenas Mais Semelhantes



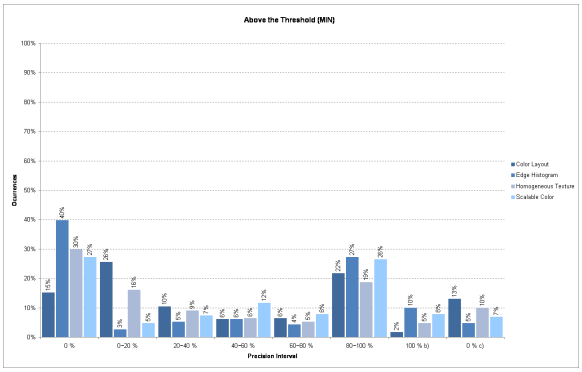
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

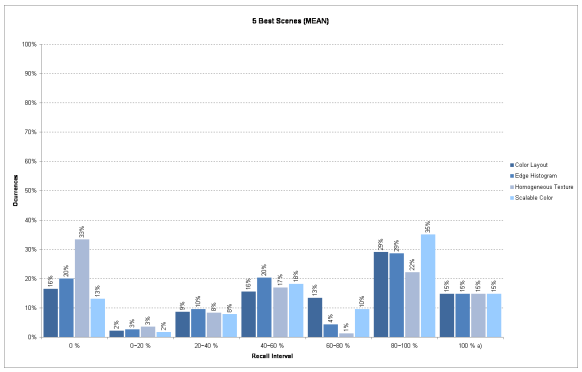


f) Cenas Semelhantes Acima do Limiar de Semelhança

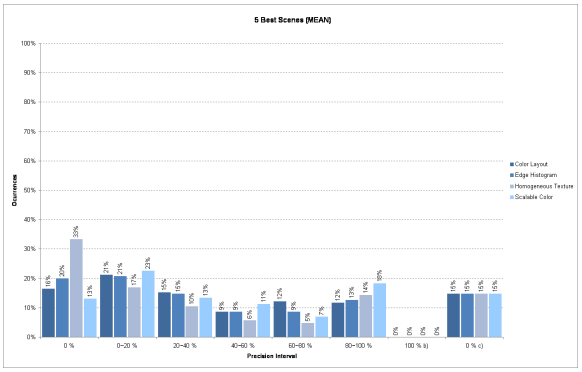
Figura C.2: Taxas de recuperação e precisão utilizando o mínimo de semelhança da cena

C.1.3 Análise utilizando a Média de semelhança da Cena

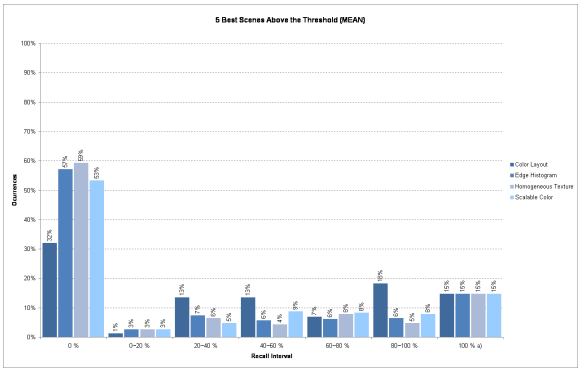
Recuperação Precisão



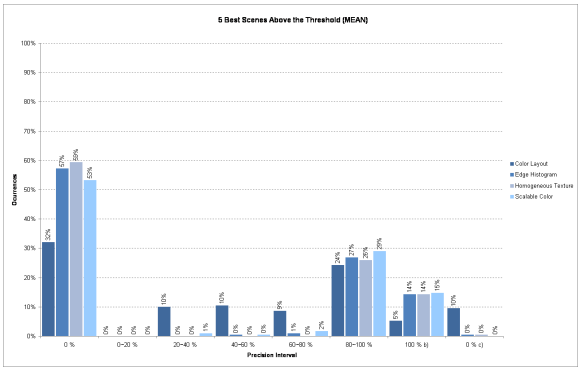
a) 5 Cenas Mais Semelhantes



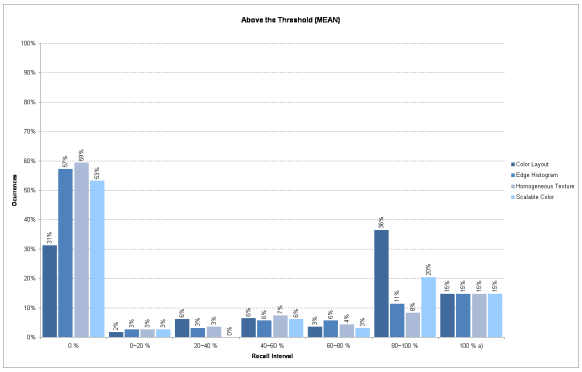
b) 5 Cenas Mais Semelhantes



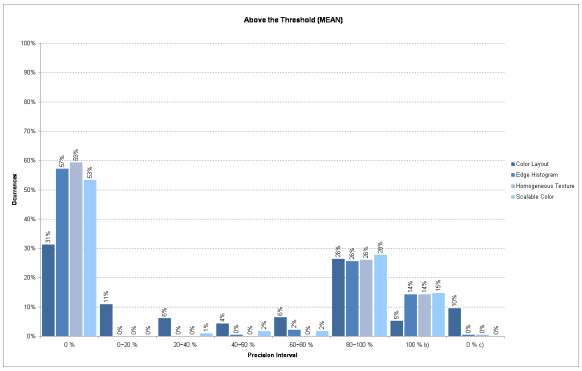
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança



f) Cenas Semelhantes Acima do Limiar de Semelhança

Figura C.3: Taxas de recuperação e precisão utilizando a média de semelhança da cena

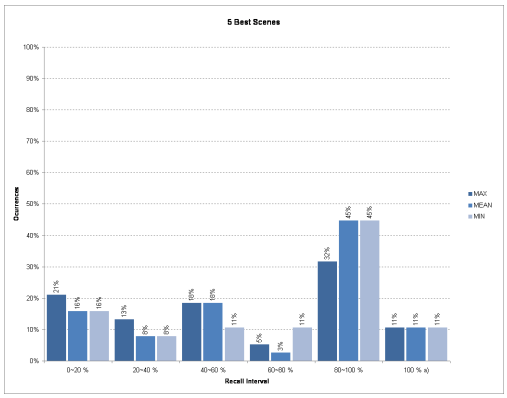
C.2 Comparativos por excerto de vídeo e tipo de Descritor

Para cada um dos excertos de vídeo *Animal* (MPEG, 1998), *Noticias TVE* (MPEG, 1998), *Concurso TVE* (MPEG, 1998), *Inspector Gadget* (imdb, 2005) e *Other Side Of Heaven* (IMDB, 2001), são apresentados os resultados para cada um dos descritores *Color Layout*, *Edge Histogram*, *Homogeneous Texture* e *Scalable Color*. Nos gráficos podem ser visualizadas as diferenças relativas à utilização dos Máximos, Mínimos ou Médias de semelhanças para representação de Cenas.

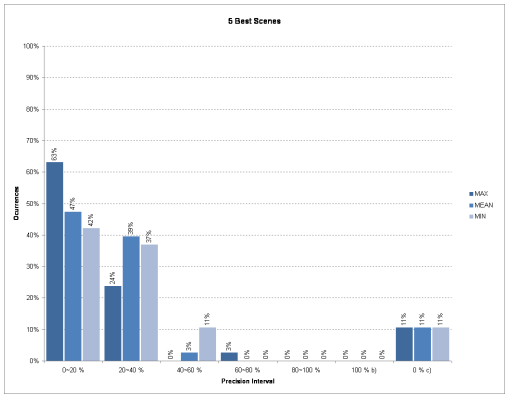
C.2.1 Excerto de vídeo *Animals*

Descritor *Color Layout*

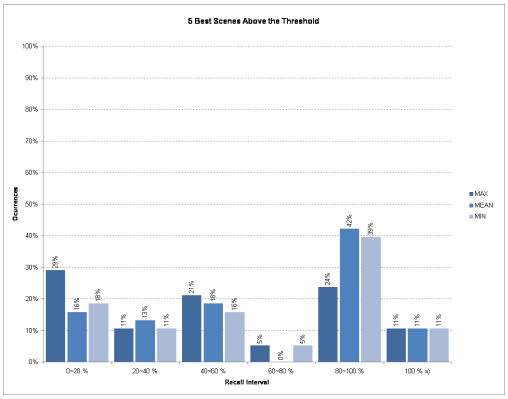
Recuperação | Precisão



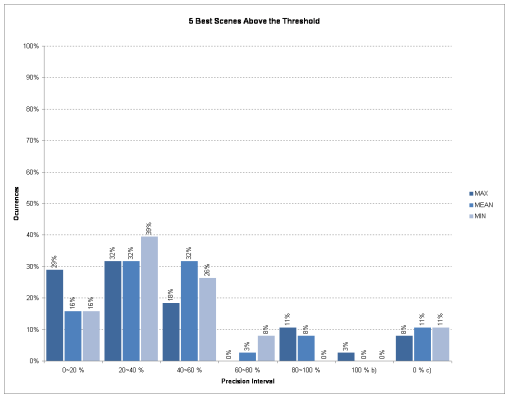
a) 5 Cenas Mais Semelhantes



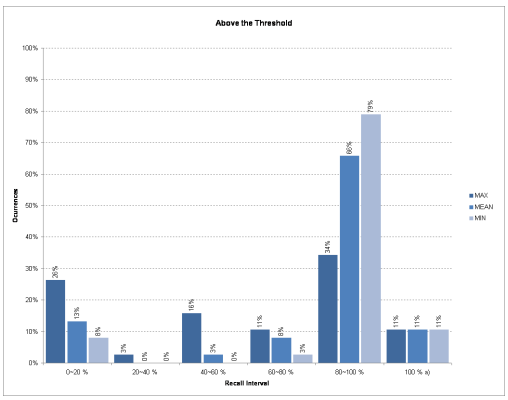
b) 5 Cenas Mais Semelhantes



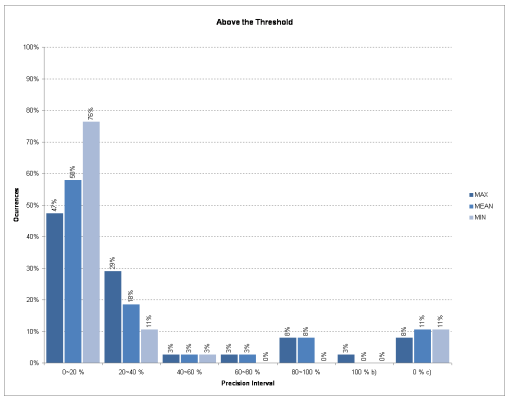
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

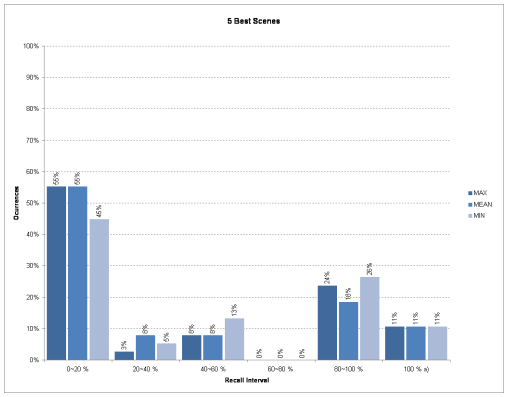


f) Cenas Semelhantes Acima do Limiar de Semelhança

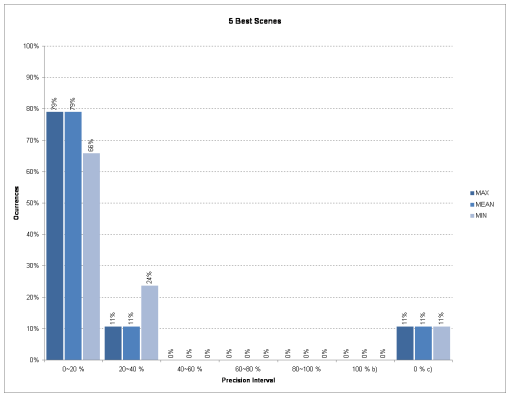
Figura C.4: Taxas de recuperação e precisão utilizando o descritor *Color Layout*

Descritor *Edge Histogram*

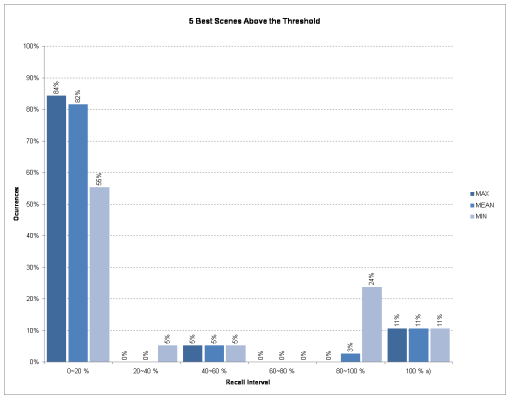
Recuperação Precisão



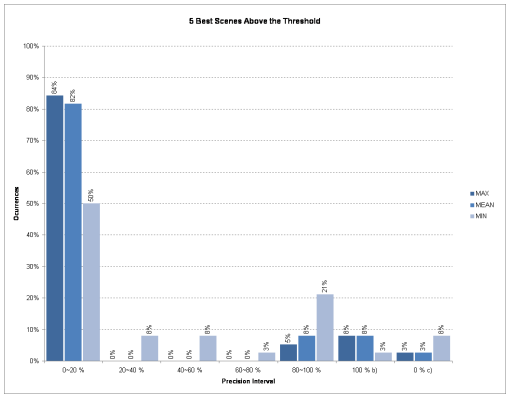
a) 5 Cenas Mais Semelhantes



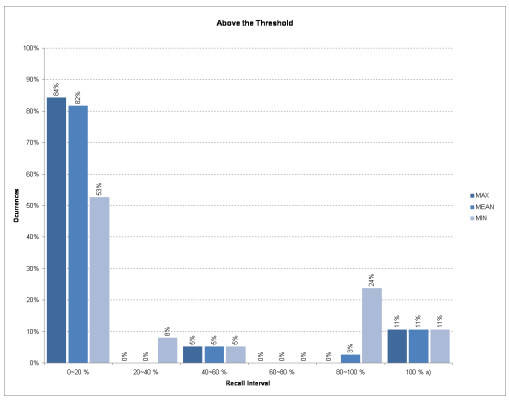
b) 5 Cenas Mais Semelhantes



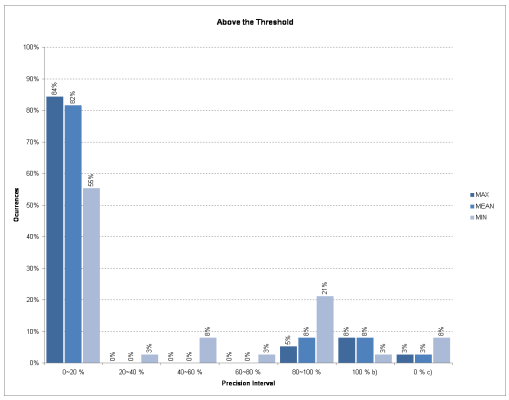
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

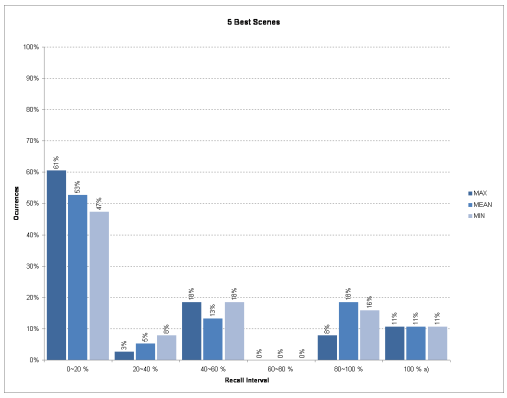


f) Cenas Semelhantes Acima do Limiar de Semelhança

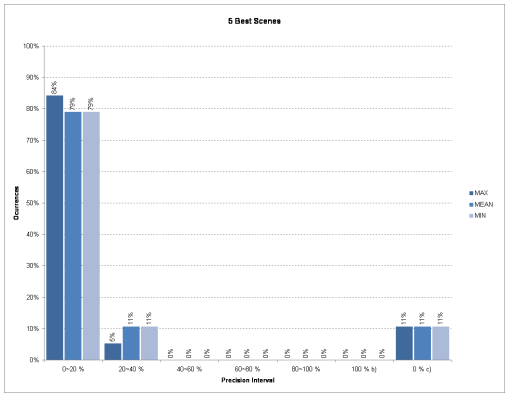
Figura C.5: Taxas de recuperação e precisão utilizando o descritor *Edge Histogram*

Descritor *Homogeneous Texture*

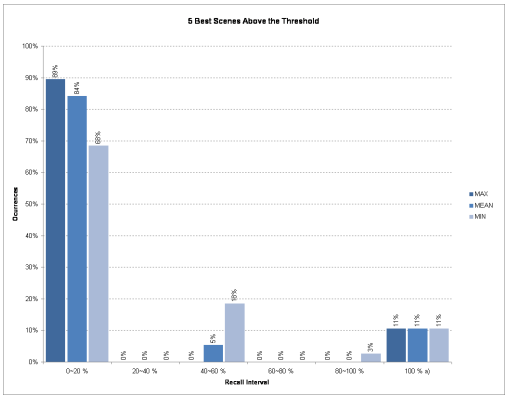
Recuperação | Precisão



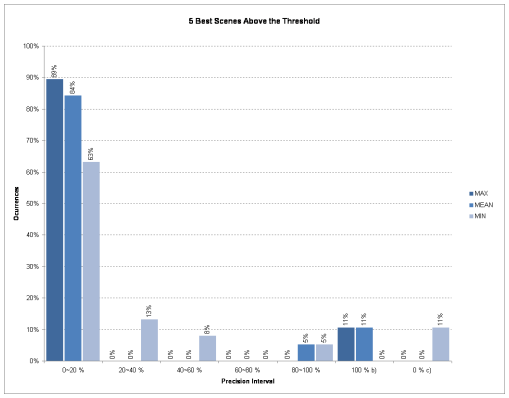
a) 5 Cenas Mais Semelhantes



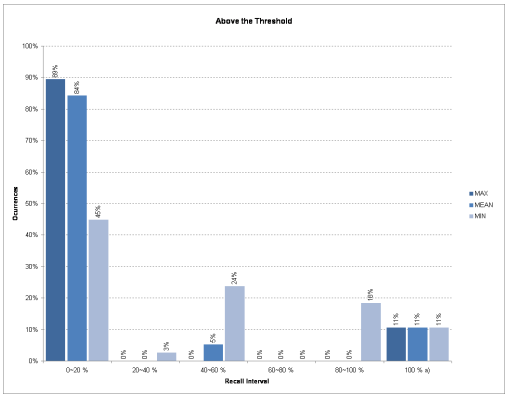
b) 5 Cenas Mais Semelhantes



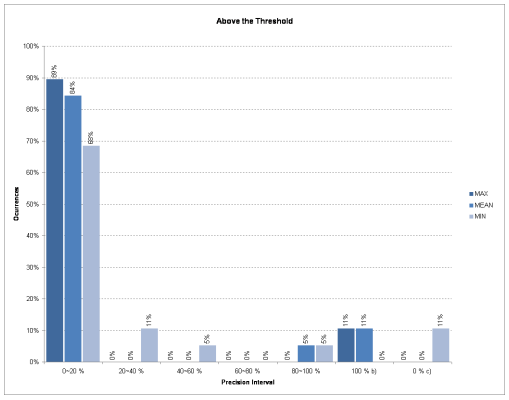
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

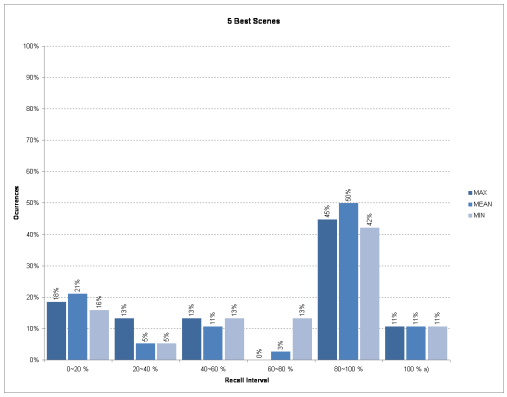


f) Cenas Semelhantes Acima do Limiar de Semelhança

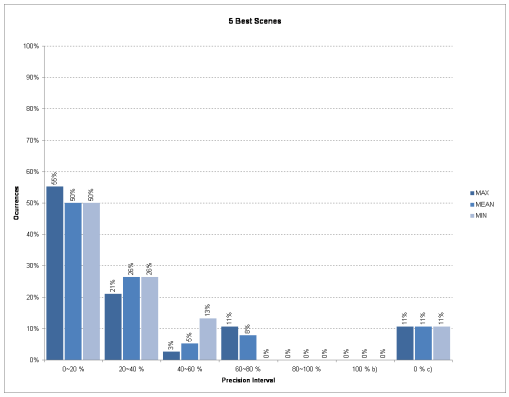
Figura C.6: Taxas de recuperação e precisão utilizando o descritor *Homogeneous Texture*

Descritor *Scalable Color*

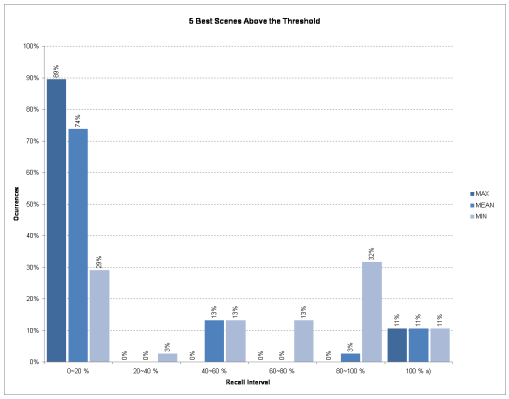
Recuperação Precisão



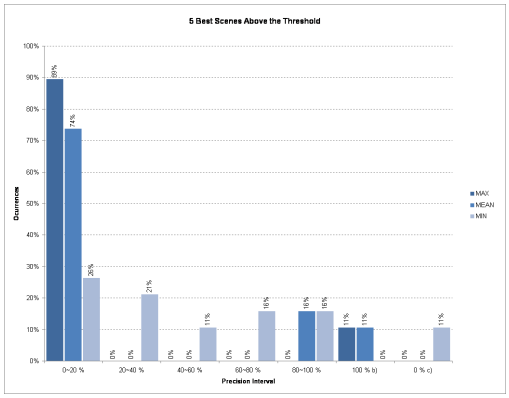
a) 5 Cenas Mais Semelhantes



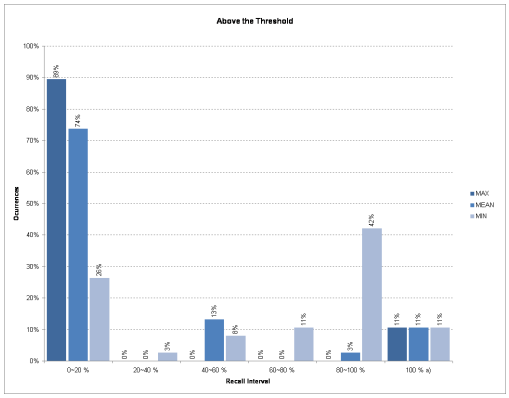
b) 5 Cenas Mais Semelhantes



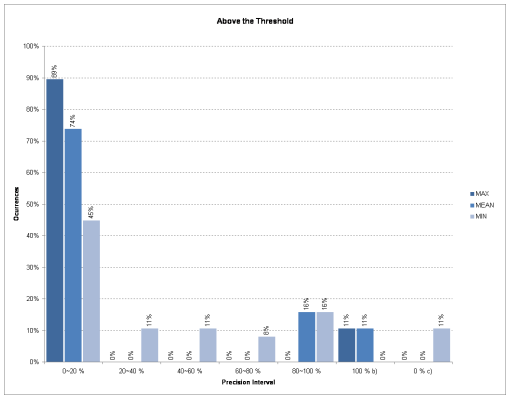
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança



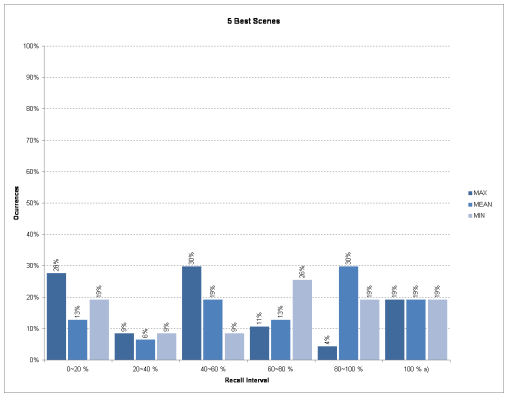
f) Cenas Semelhantes Acima do Limiar de Semelhança

Figura C.7: Taxas de recuperação e precisão utilizando o descritor *Scalable Color*

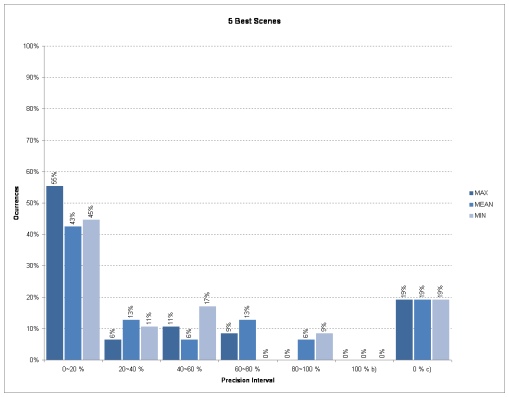
C.2.2 Excerto de vídeo *Noticias TVE*

Descritor *Color Layout*

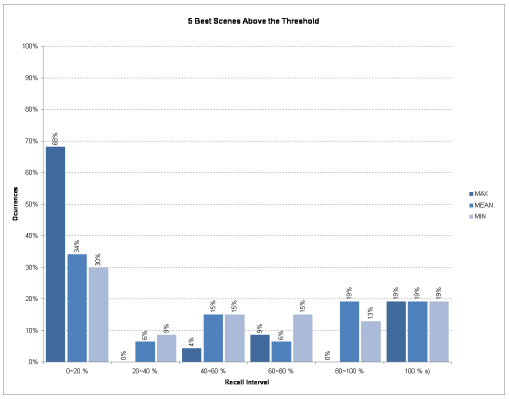
Recuperação Precisão



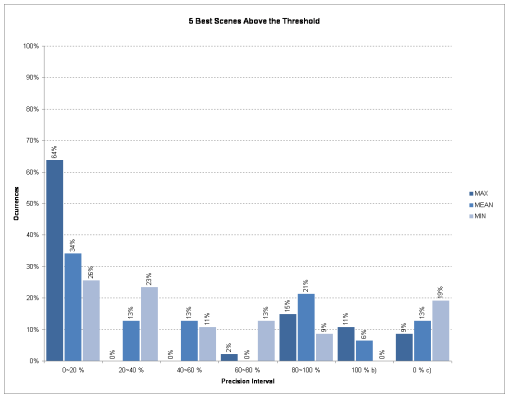
a) 5 Cenas Mais Semelhantes



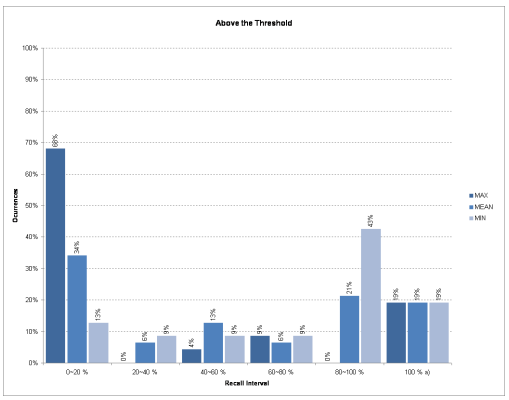
b) 5 Cenas Mais Semelhantes



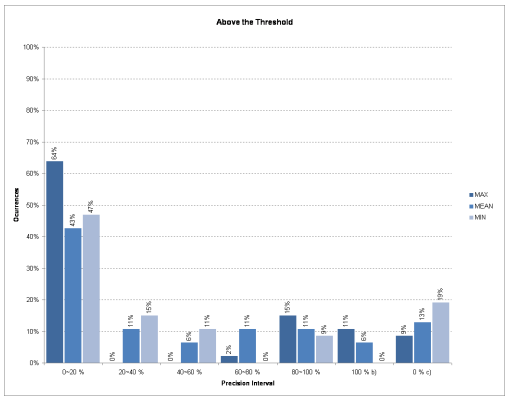
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

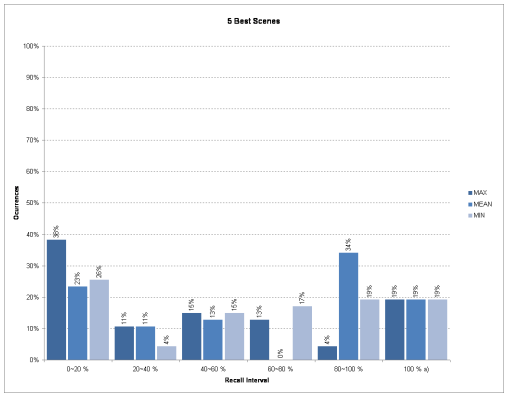


f) Cenas Semelhantes Acima do Limiar de Semelhança

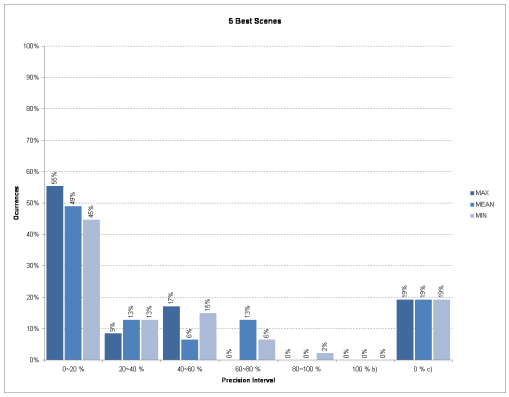
Figura C.8: Taxas de recuperação e precisão utilizando o descritor *Color Layout*

Descritor *Edge Histogram*

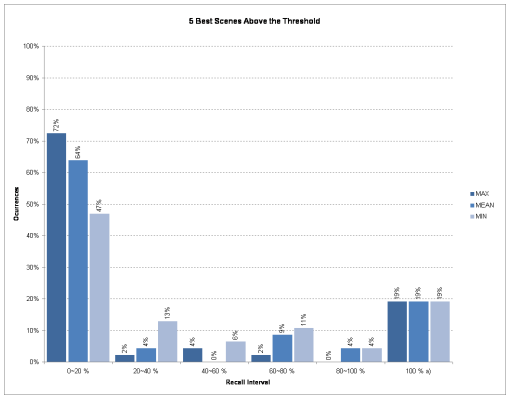
Recuperação Precisão



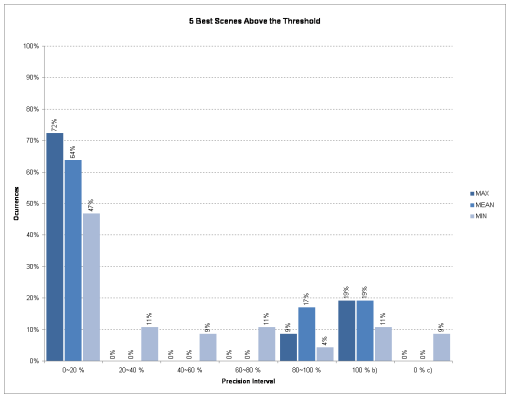
a) 5 Cenas Mais Semelhantes



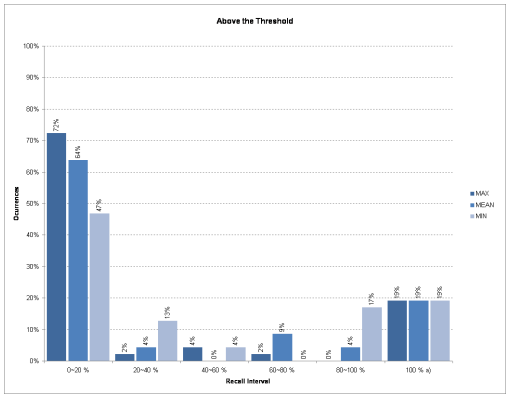
b) 5 Cenas Mais Semelhantes



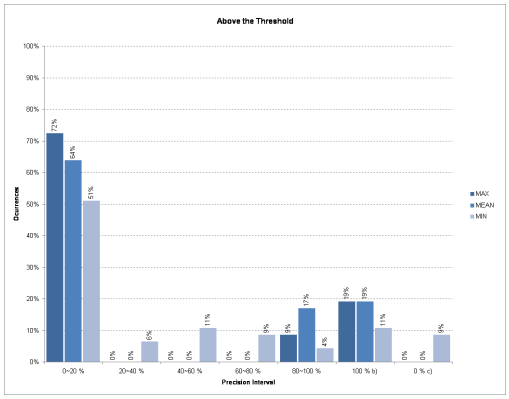
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

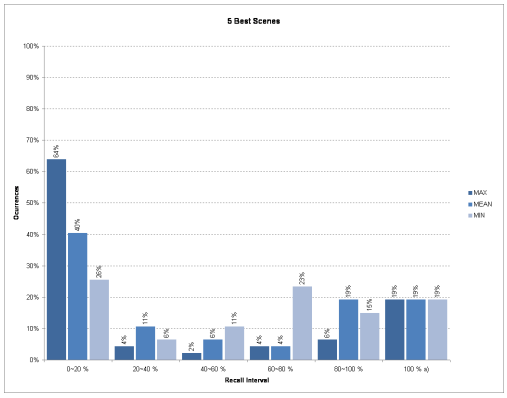


f) Cenas Semelhantes Acima do Limiar de Semelhança

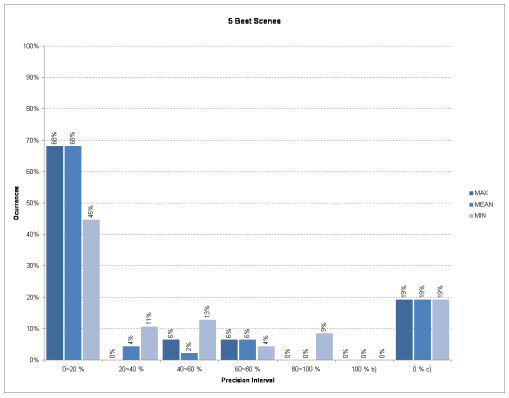
Figura C.9: Taxas de recuperação e precisão utilizando o descritor *Edge Histogram*

Descritor *Homogeneous Texture*

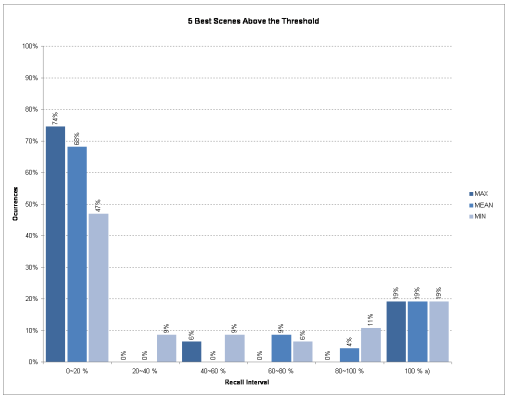
Recuperação Precisão



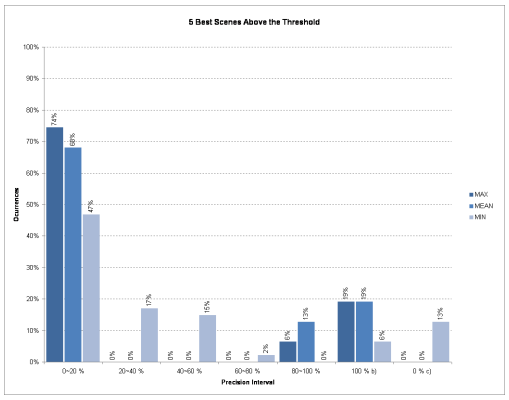
a) 5 Cenas Mais Semelhantes



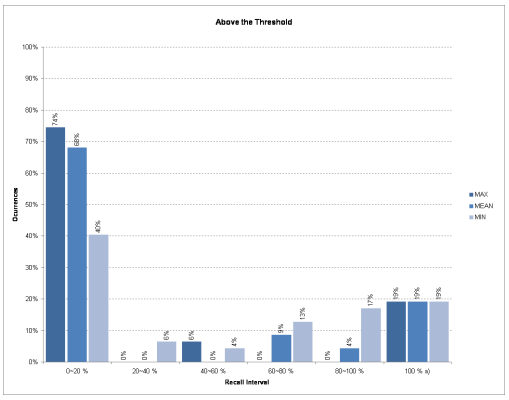
b) 5 Cenas Mais Semelhantes



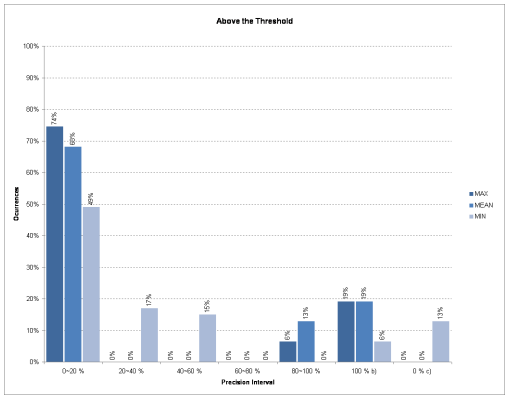
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

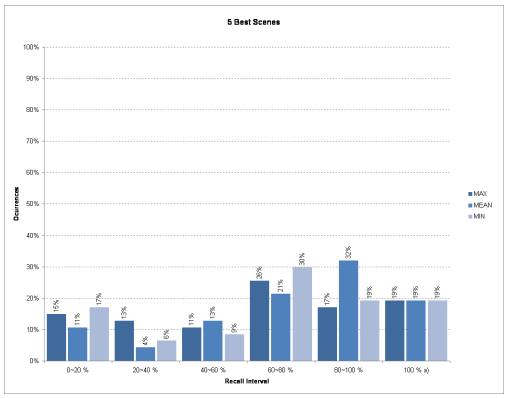


f) Cenas Semelhantes Acima do Limiar de Semelhança

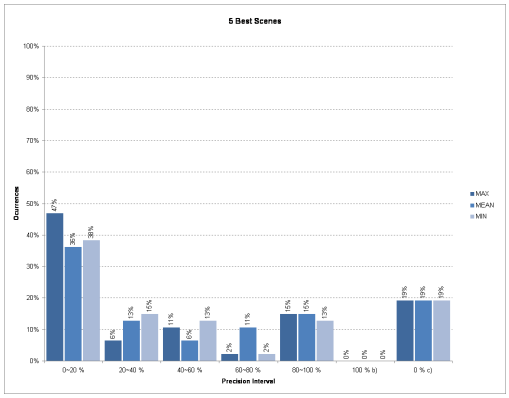
Figura C.10: Taxas de recuperação e precisão utilizando o descritor *Homogeneous Texture*

Descritor *Scalable Color*

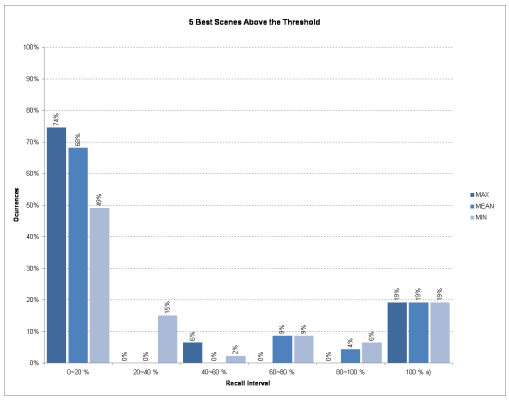
Recuperação | Precisão



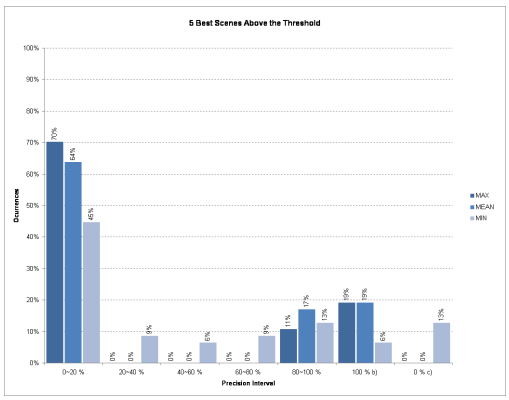
a) 5 Cenas Mais Semelhantes



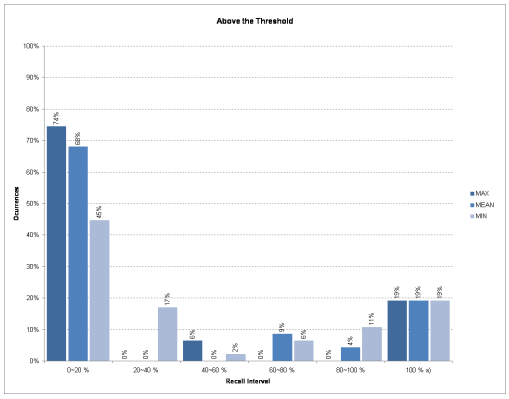
b) 5 Cenas Mais Semelhantes



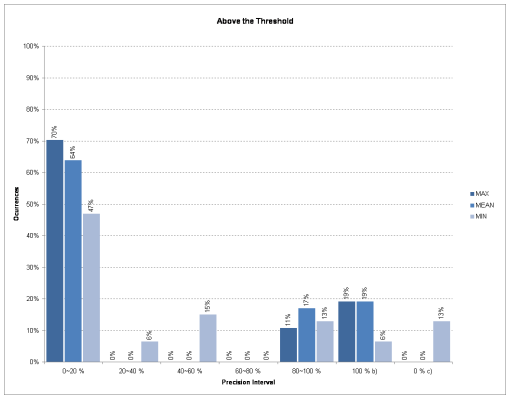
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança



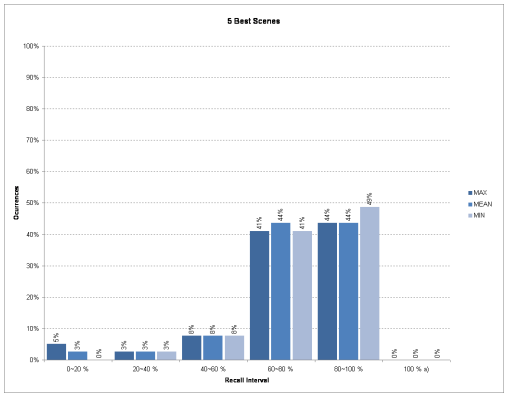
f) Cenas Semelhantes Acima do Limiar de Semelhança

Figura C.11: Taxas de recuperação e precisão utilizando o descritor *Scalable Color*

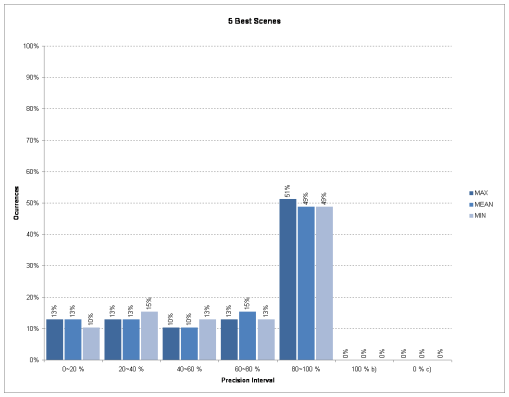
C.2.3 Excerto de vídeo *Concurso TVE*

Descritor *Color Layout*

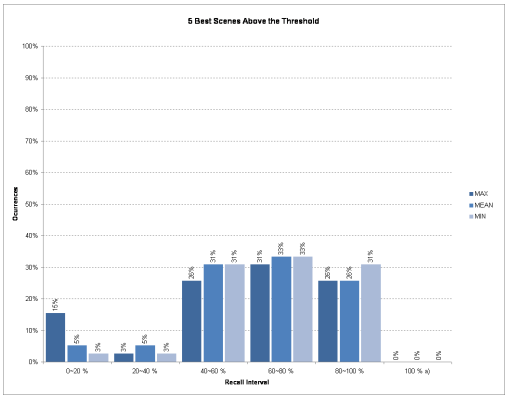
Recuperação Precisão



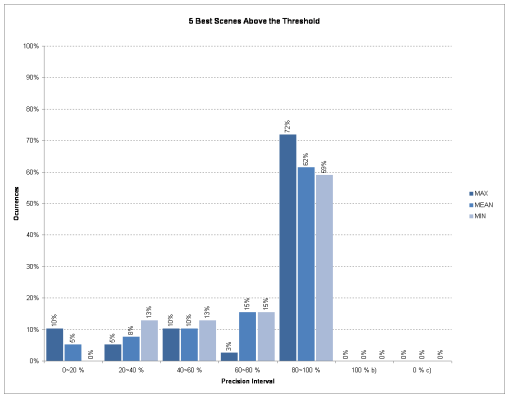
a) 5 Cenas Mais Semelhantes



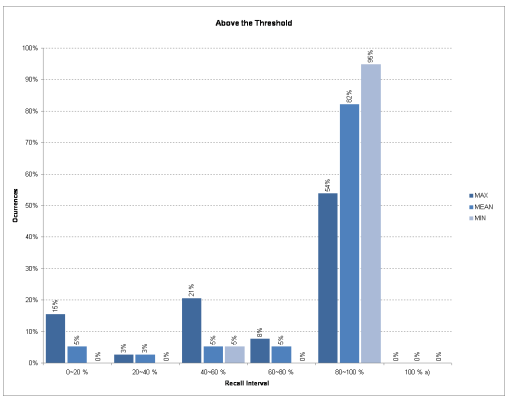
b) 5 Cenas Mais Semelhantes



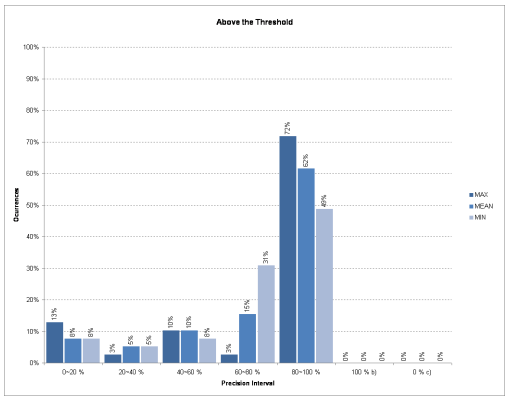
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

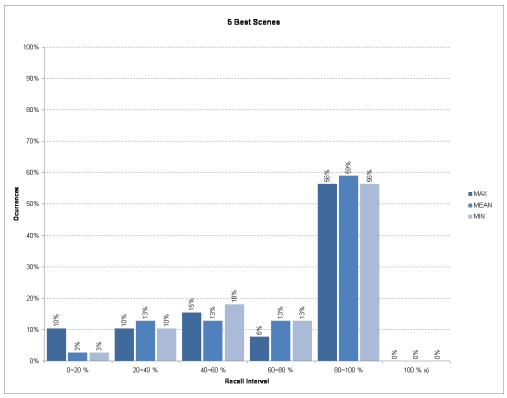


f) Cenas Semelhantes Acima do Limiar de Semelhança

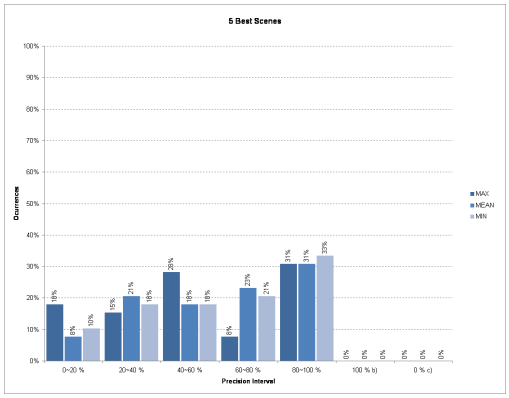
Figura C.12: Taxas de recuperação e precisão utilizando o descritor *Color Layout*

Descritor *Edge Histogram*

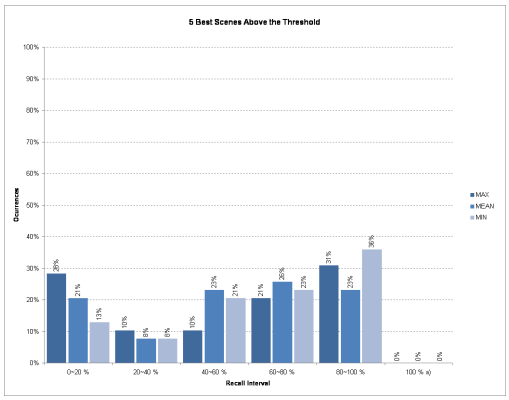
Recuperação | Precisão



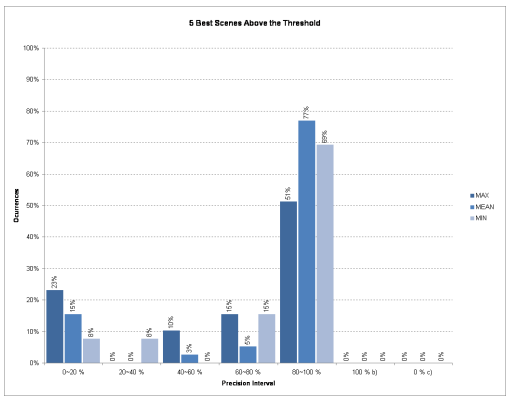
a) 5 Cenas Mais Semelhantes



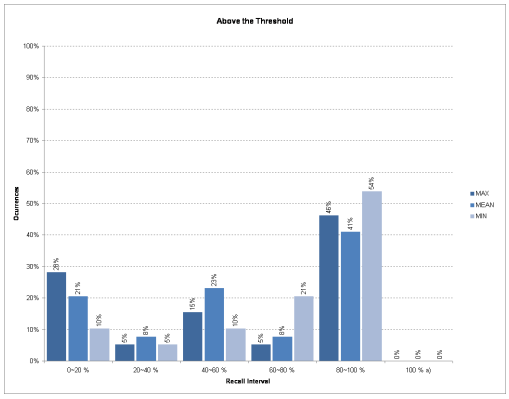
b) 5 Cenas Mais Semelhantes



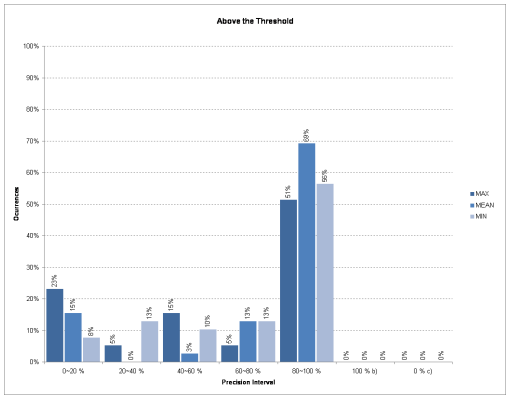
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

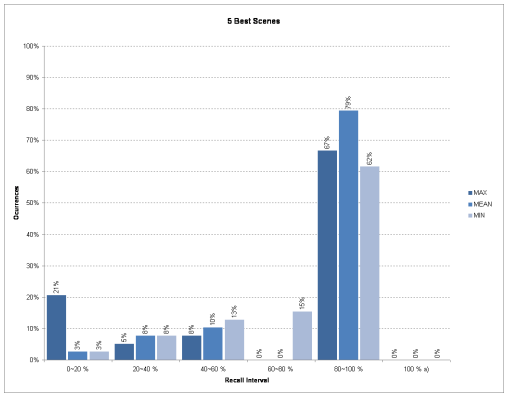


f) Cenas Semelhantes Acima do Limiar de Semelhança

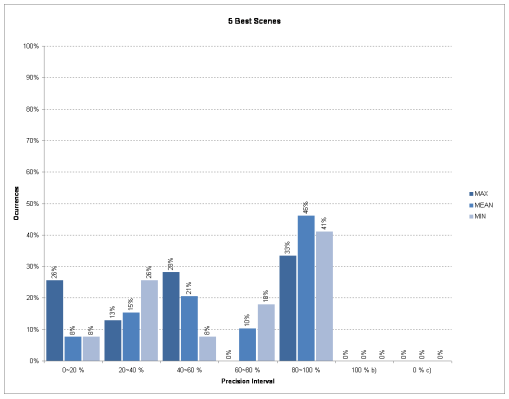
Figura C.13: Taxas de recuperação e precisão utilizando o descritor *Edge Histogram*

Descritor *Homogeneous Texture*

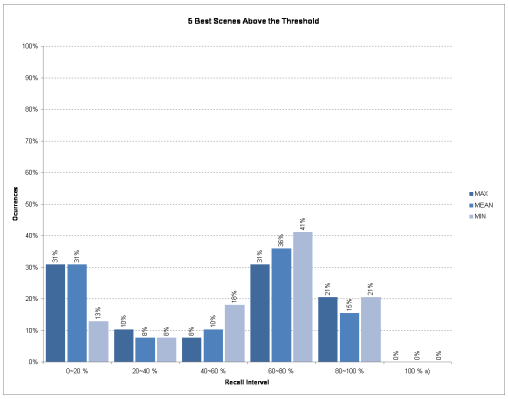
Recuperação Precisão



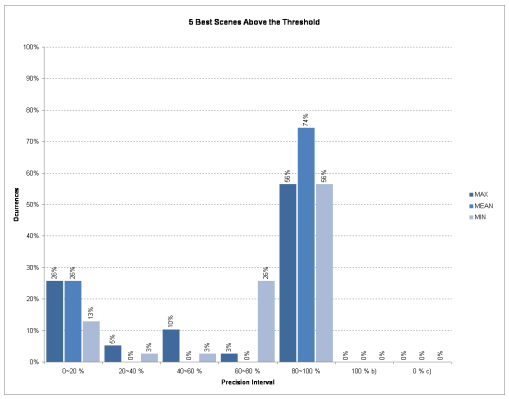
a) 5 Cenas Mais Semelhantes



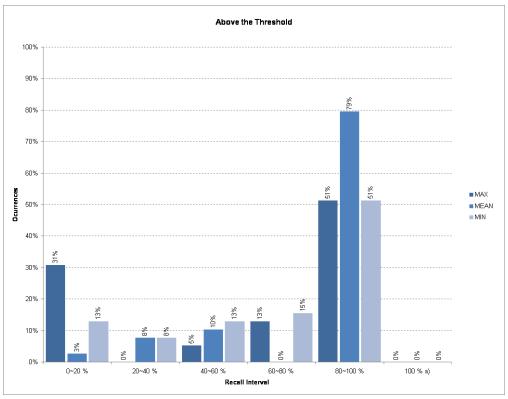
b) 5 Cenas Mais Semelhantes



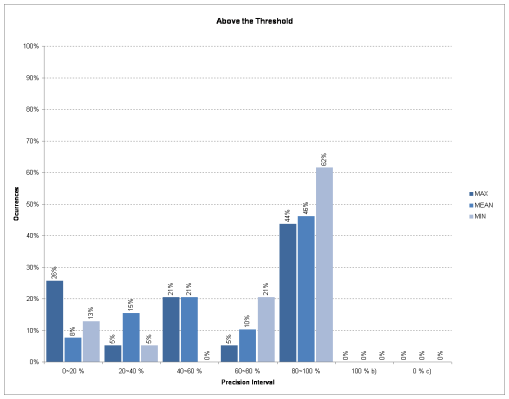
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

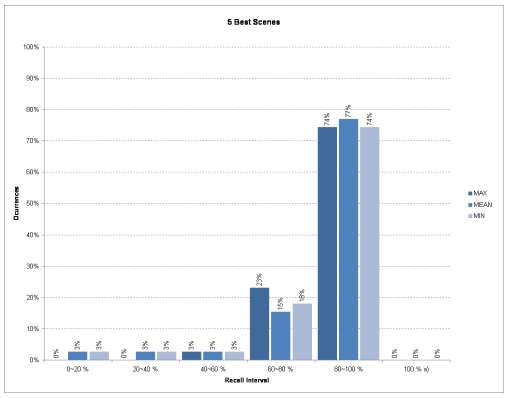


f) Cenas Semelhantes Acima do Limiar de Semelhança

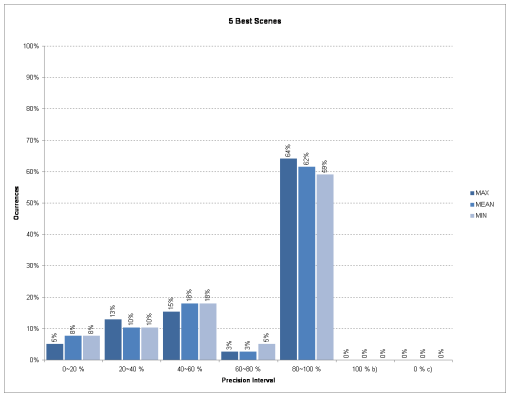
Figura C.14: Taxas de recuperação e precisão utilizando o descritor *Homogeneous Texture*

Descritor *Scalable Color*

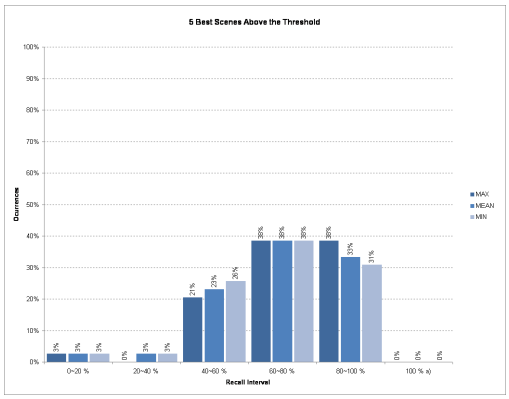
Recuperação | Precisão



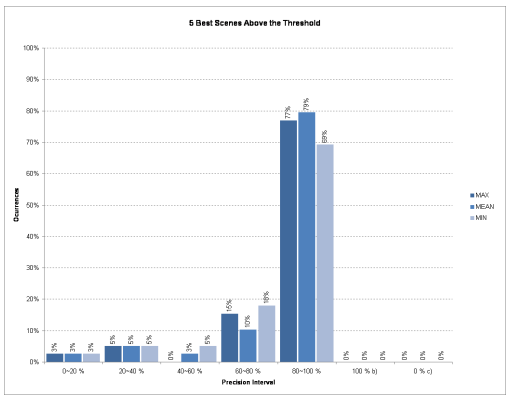
a) 5 Cenas Mais Semelhantes



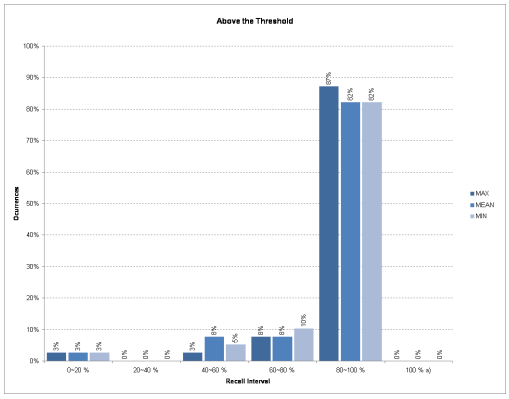
b) 5 Cenas Mais Semelhantes



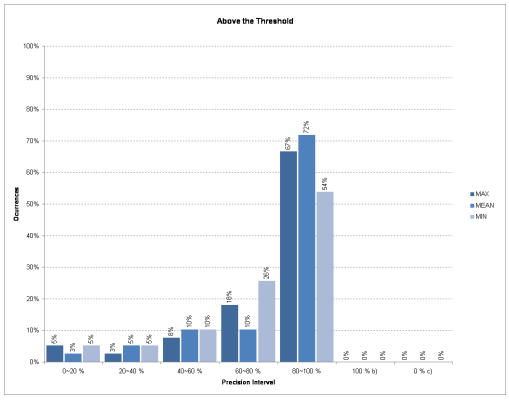
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança



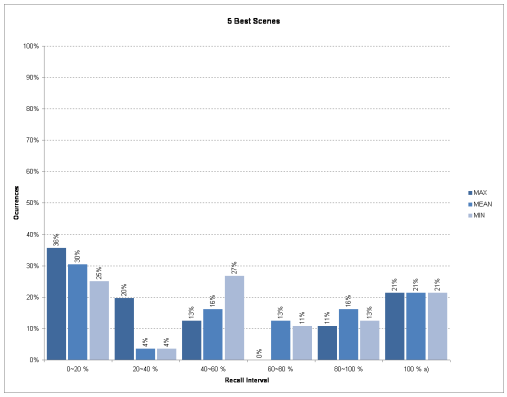
f) Cenas Semelhantes Acima do Limiar de Semelhança

Figura C.15: Taxas de recuperação e precisão utilizando o descritor *Scalable Color*

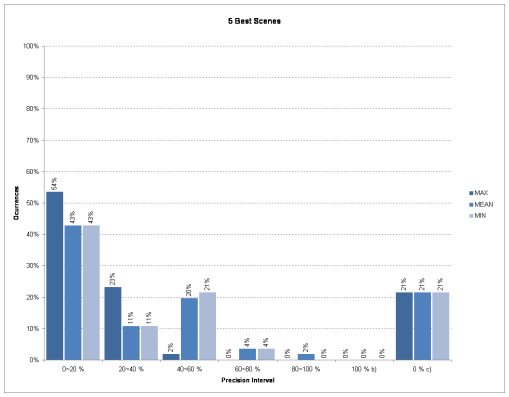
C.2.4 Excerto de vídeo *Inspector Gadget*

Descritor *Color Layout*

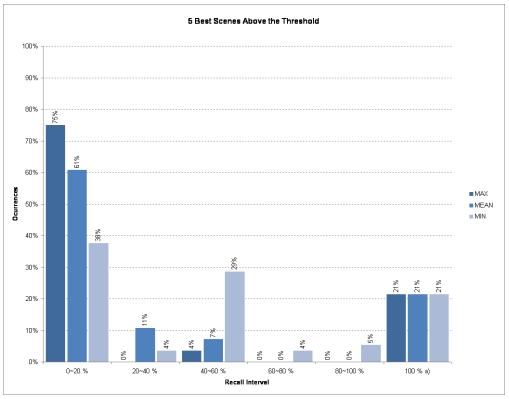
Recuperação Precisão



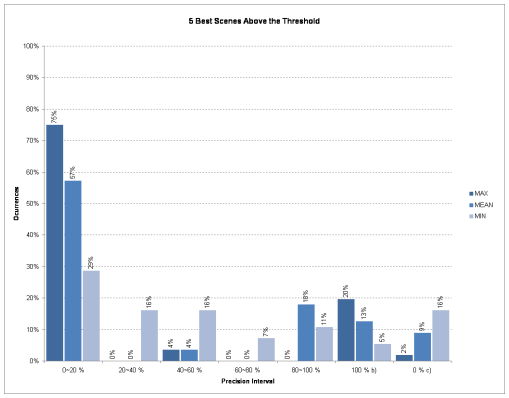
a) 5 Cenas Mais Semelhantes



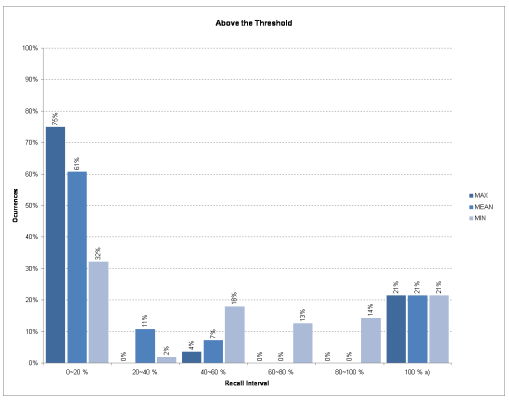
b) 5 Cenas Mais Semelhantes



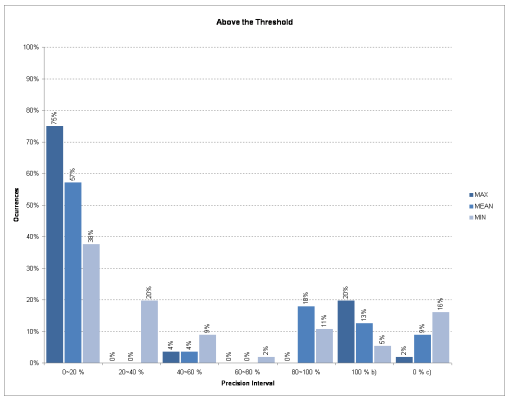
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

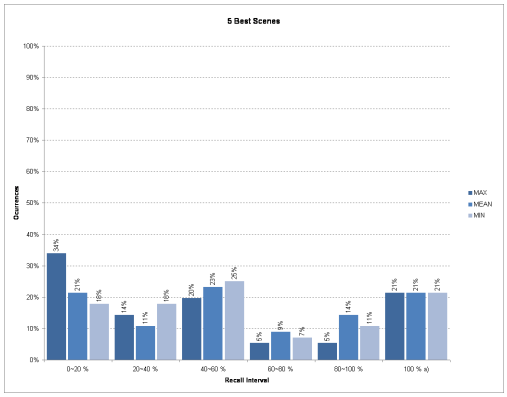


f) Cenas Semelhantes Acima do Limiar de Semelhança

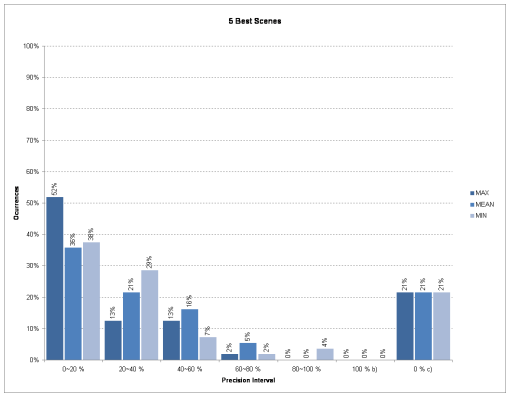
Figura C.16: Taxas de recuperação e precisão utilizando o descritor *Color Layout*

Descritor *Edge Histogram*

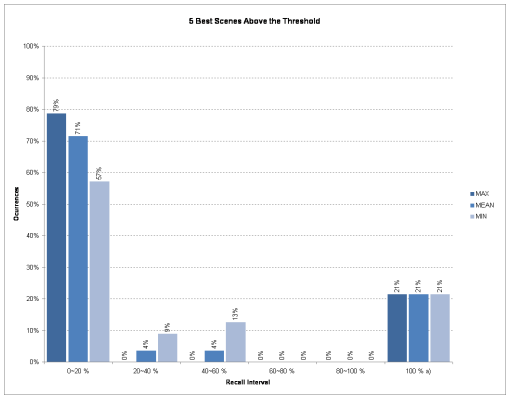
Recuperação Precisão



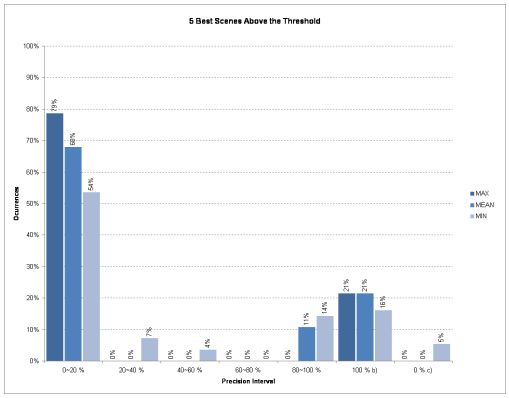
a) 5 Cenas Mais Semelhantes



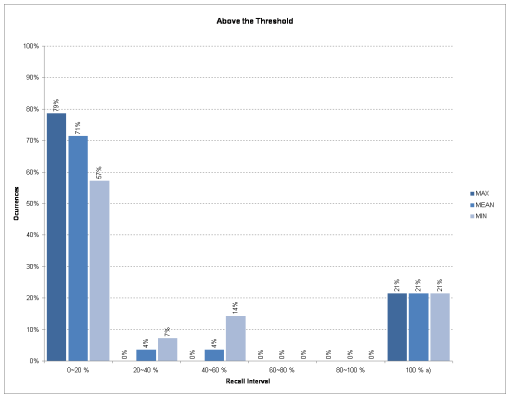
b) 5 Cenas Mais Semelhantes



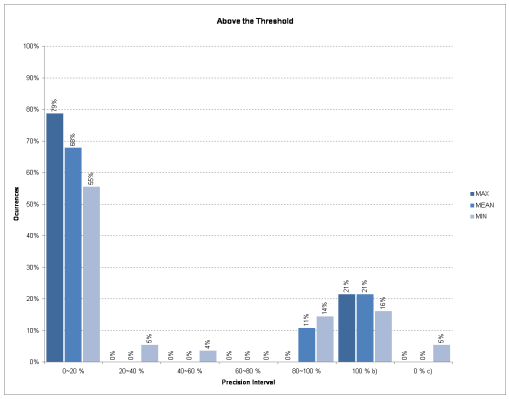
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

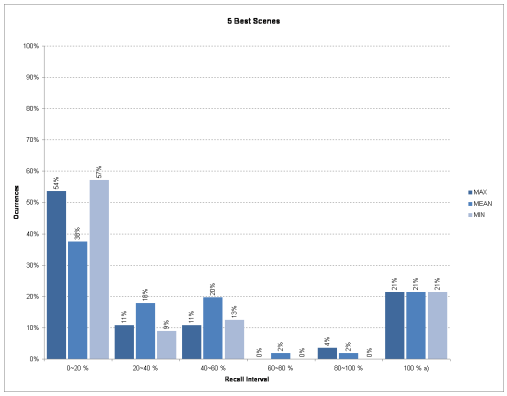


f) Cenas Semelhantes Acima do Limiar de Semelhança

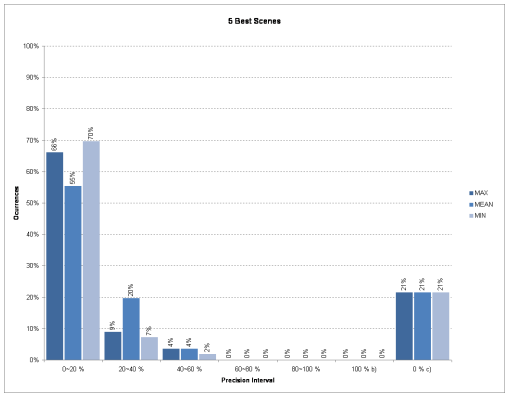
Figura C.17: Taxas de recuperação e precisão utilizando o descritor *Edge Histogram*

Descritor *Homogeneous Texture*

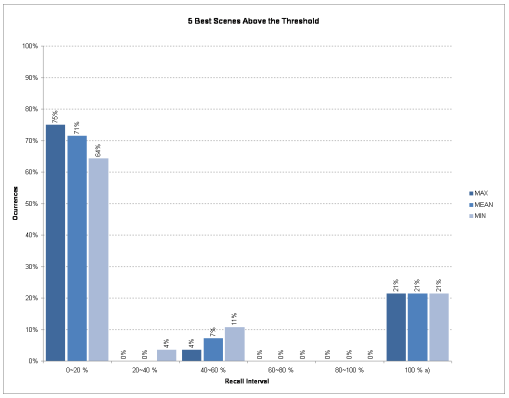
Recuperação Precisão



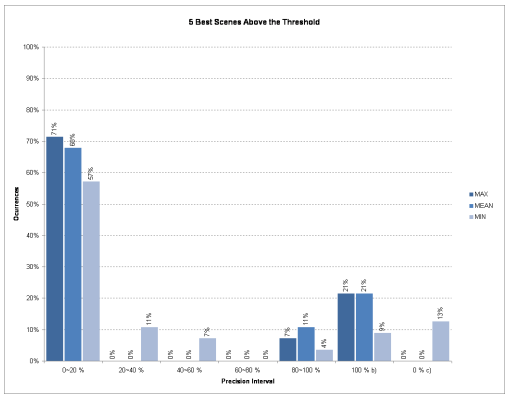
a) 5 Cenas Mais Semelhantes



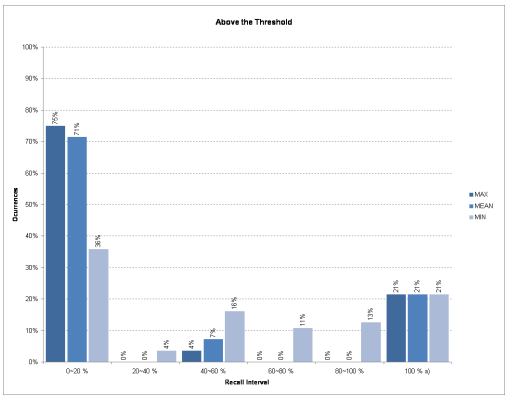
b) 5 Cenas Mais Semelhantes



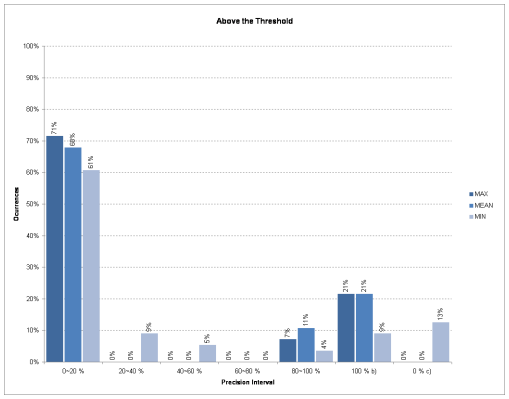
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

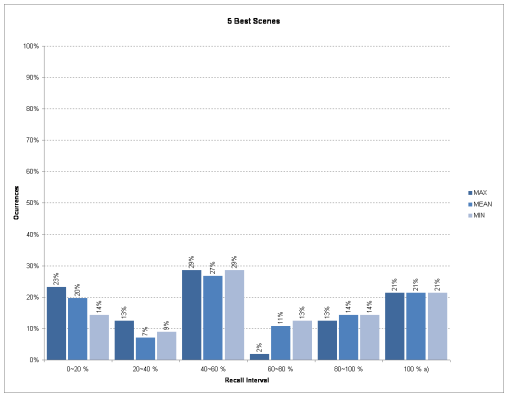


f) Cenas Semelhantes Acima do Limiar de Semelhança

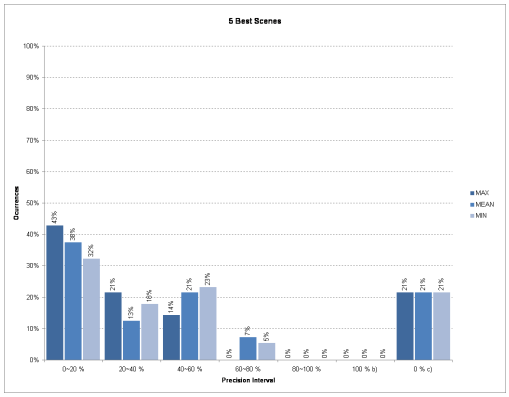
Figura C.18: Taxas de recuperação e precisão utilizando o descritor *Homogeneous Texture*

Descritor *Scalable Color*

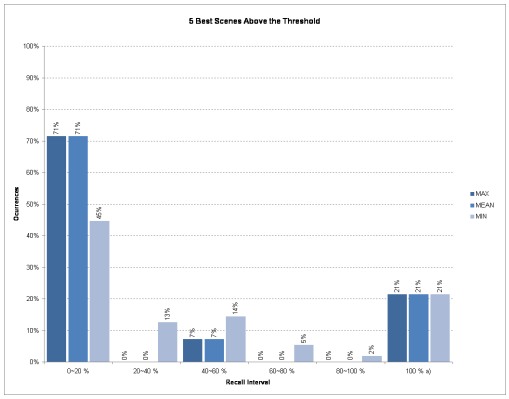
Recuperação | Precisão



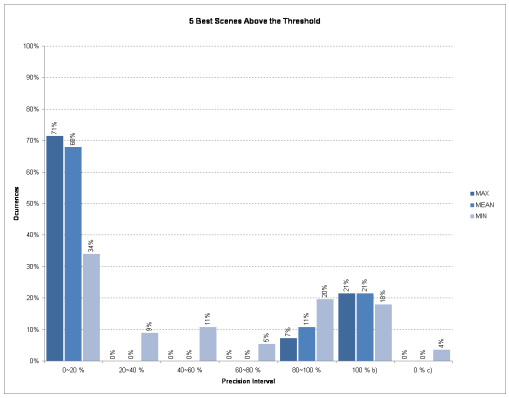
a) 5 Cenas Mais Semelhantes



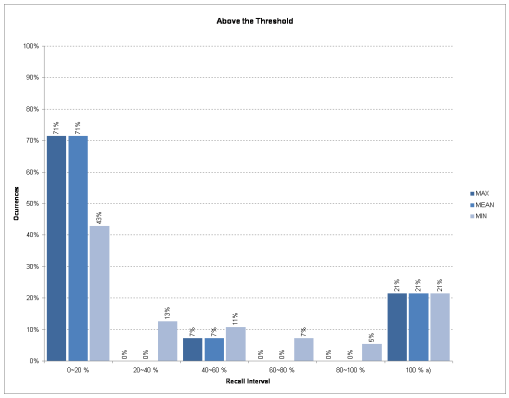
b) 5 Cenas Mais Semelhantes



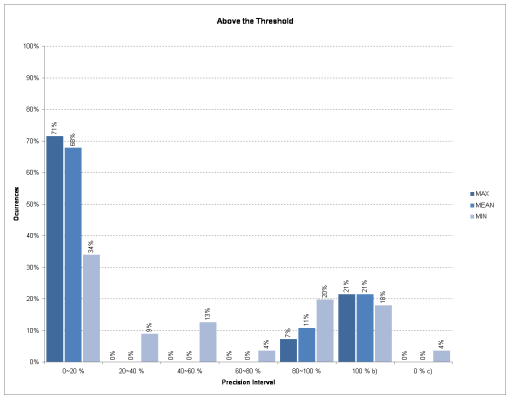
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança



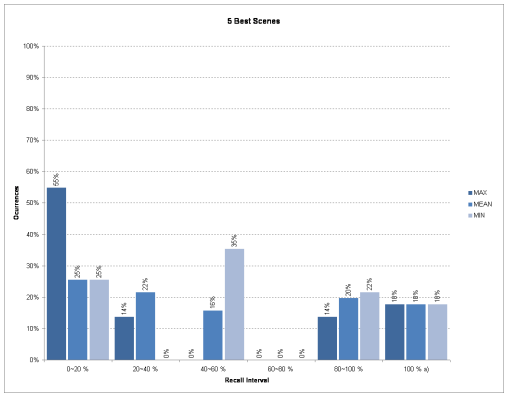
f) Cenas Semelhantes Acima do Limiar de Semelhança

Figura C.19: Taxas de recuperação e precisão utilizando o descritor *Scalable Color*

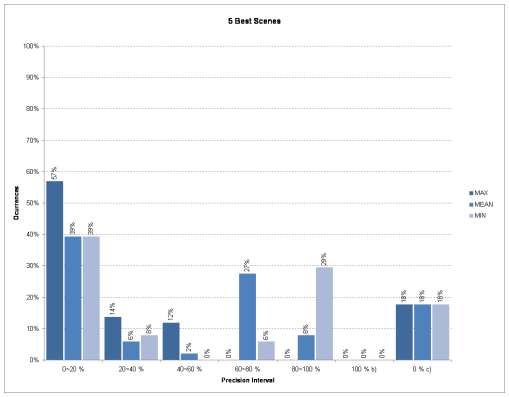
C.2.5 Excerto de vídeo *Other Side Of Heaven*

Descritor *Color Layout*

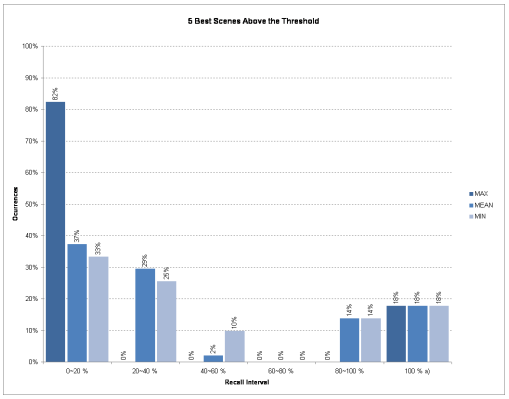
Recuperação Precisão



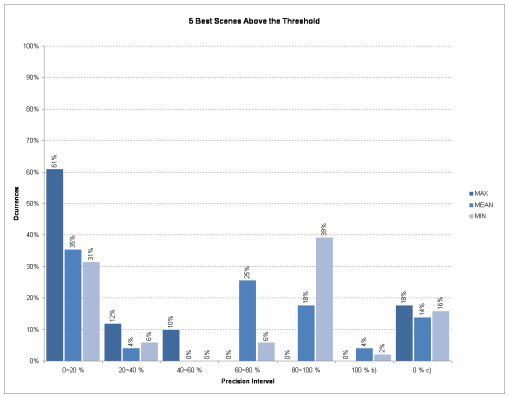
a) 5 Cenas Mais Semelhantes



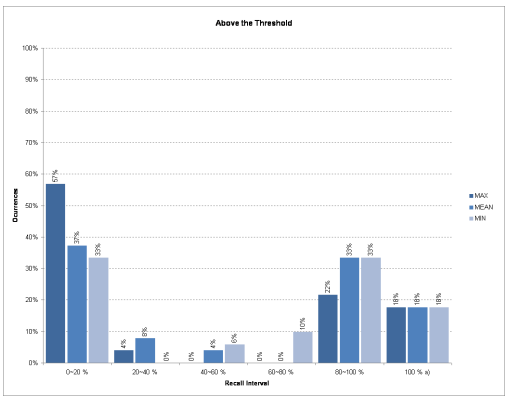
b) 5 Cenas Mais Semelhantes



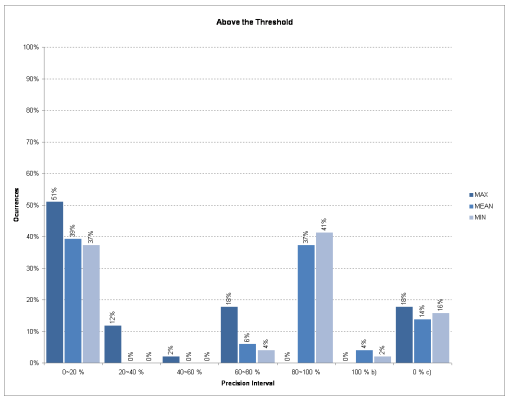
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

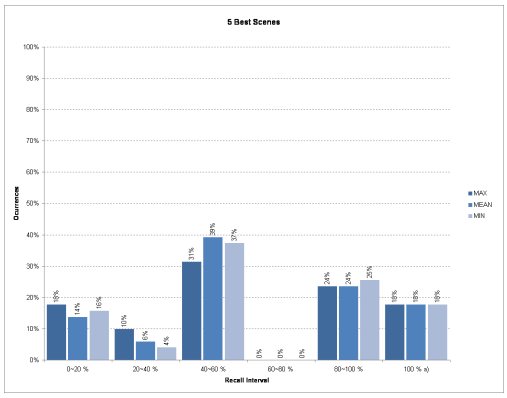


f) Cenas Semelhantes Acima do Limiar de Semelhança

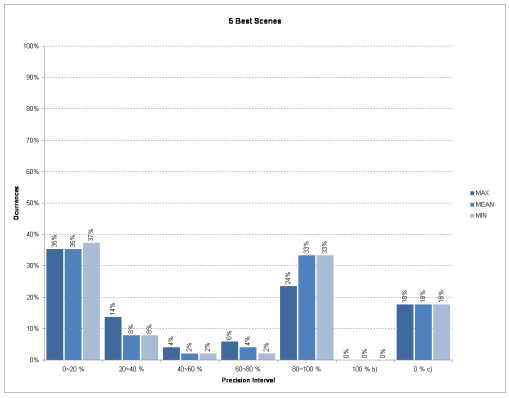
Figura C.20: Taxas de recuperação e precisão utilizando o descritor *Color Layout*

Descritor *Edge Histogram*

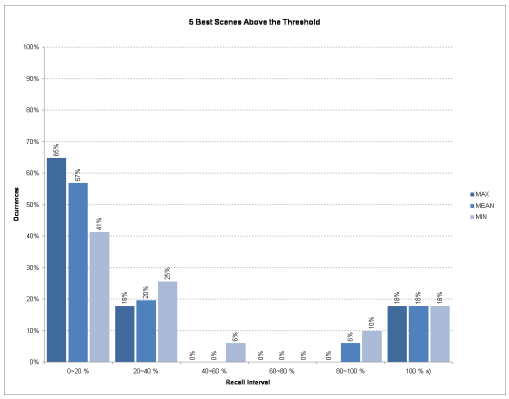
Recuperação | Precisão



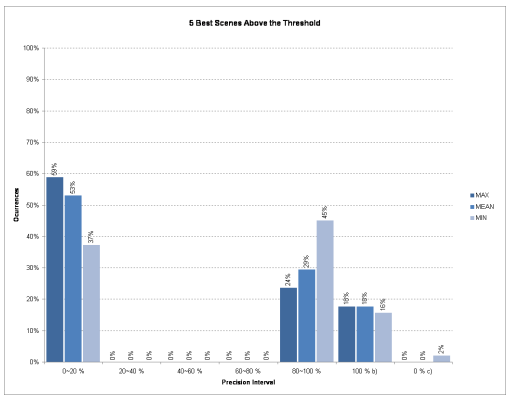
a) 5 Cenas Mais Semelhantes



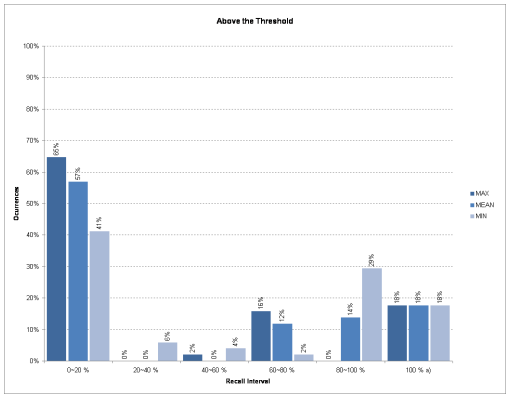
b) 5 Cenas Mais Semelhantes



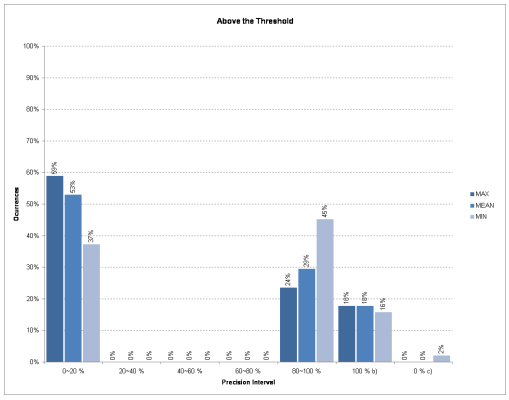
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

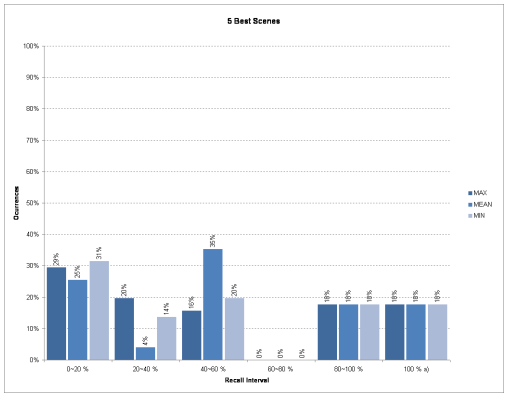


f) Cenas Semelhantes Acima do Limiar de Semelhança

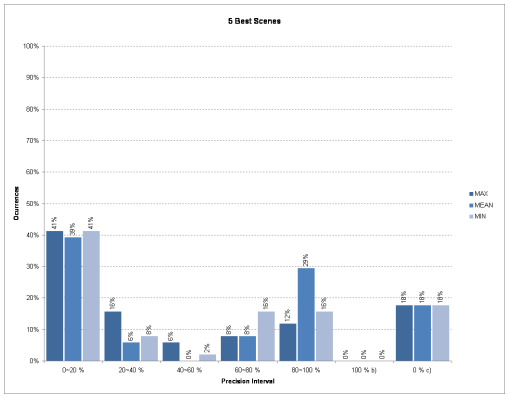
Figura C.21: Taxas de recuperação e precisão utilizando o descritor *Edge Histogram*

Descritor *Homogeneous Texture*

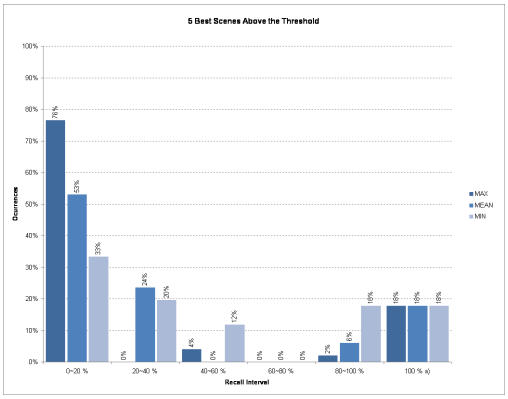
Recuperação | Precisão



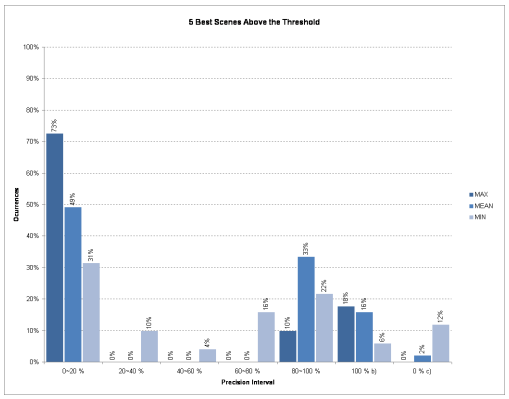
a) 5 Cenas Mais Semelhantes



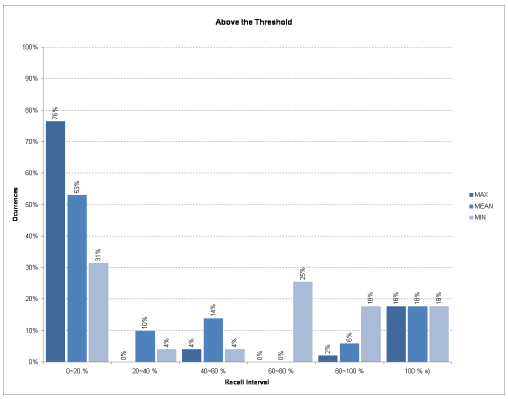
b) 5 Cenas Mais Semelhantes



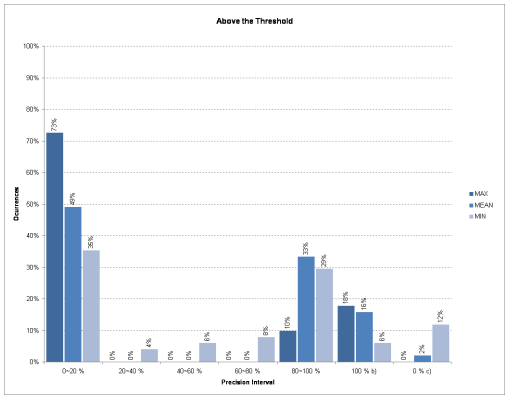
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança

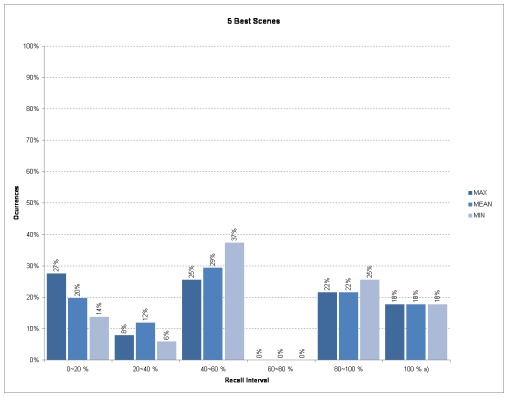


f) Cenas Semelhantes Acima do Limiar de Semelhança

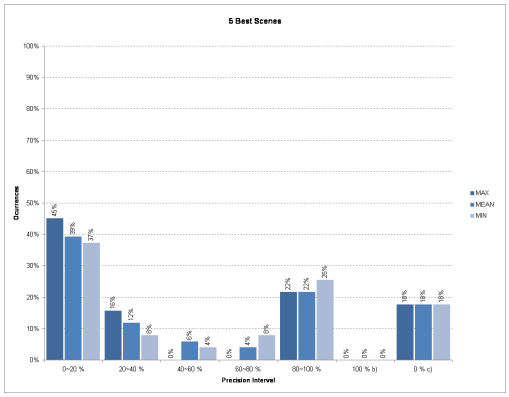
Figura C.22: Taxas de recuperação e precisão utilizando o descritor *Homogeneous Texture*

Descritor *Scalable Color*

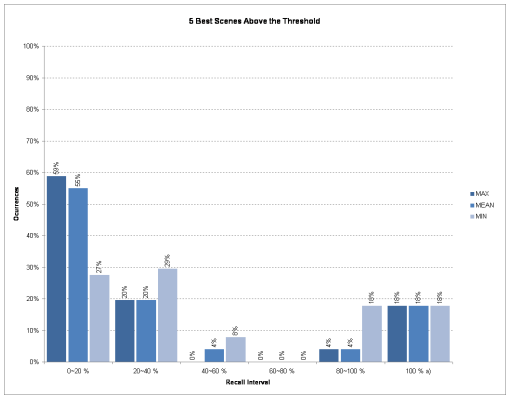
Recuperação | Precisão



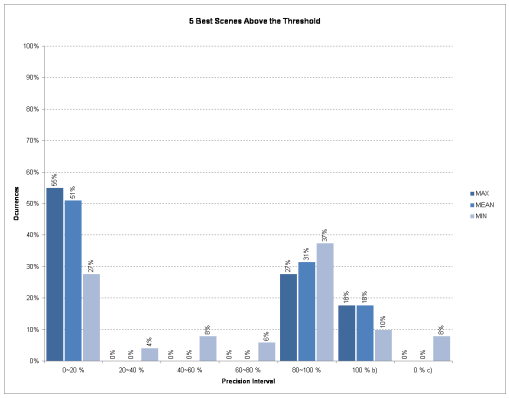
a) 5 Cenas Mais Semelhantes



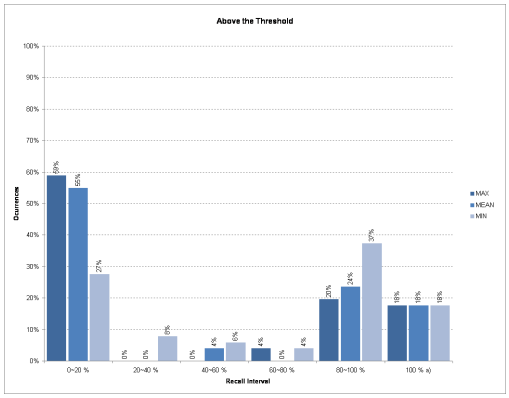
b) 5 Cenas Mais Semelhantes



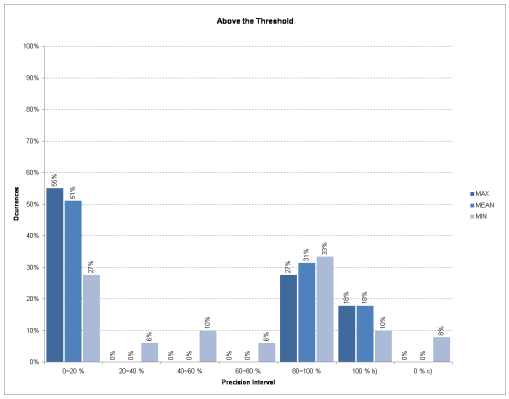
c) 5 Cenas Mais Semelhantes Acima do Limiar de Semelhança



d) 5 Cenas Mais Semelhante Acima do Limiar de Semelhança



e) Cenas Semelhantes Acima do Limiar de Semelhança



f) Cenas Semelhantes Acima do Limiar de Semelhança

Figura C.23: Taxas de recuperação e precisão utilizando o descritor *Scalable Color*