Extração de Quadros-Chave Como Subsídio Para Personalização em Vídeos Digitais

Gilson Araújo do Nascimento
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Av. Trabalhador São Carlense, 400
PO Box 668 – 13560-970
São Carlos, SP – Brasil
gilaraujo@grad.icmc.usp.br

Marcelo Garcia Manzato
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Av. Trabalhador São Carlense, 400
PO Box 668 – 13560-970
São Carlos, SP – Brasil
mmanzato@icmc.usp.br

Rudinei Goularte
Instituto de Ciências Matemáticas e de Computação
Universidade de São Paulo
Av. Trabalhador São Carlense, 400
PO Box 668 – 13560-970
São Carlos, SP – Brasil
rudinei@icmc.usp.br

Resumo

Este trabalho apresenta uma análise de técnicas de extração de quadros-chave em vídeos digitais no contexto de adaptação e personalização de conteúdo, com o propósito de facilitar a obtenção de informações semânticas de vídeos mediante detecção de faces. Isso implica em minimizar o custo computacional necessário, sem comprometer a eficiência da obtenção dessas informações através da detecção facial.

1. INTRODUÇÃO

Com o acesso à Internet, é possível enviar e receber informações de qualquer parte do mundo. Constantes avanços possibilitaram manter-se em contato com pessoas utilizando até mesmo conversas com áudio e vídeo.

Graças à utilização da banda larga, é possível transmitir grande quantidade de dados em um intervalo de tempo muito menor e os usuários tem cada vez mais interatividade e liberdade no acesso ao conteúdo digital.

Além disso, o acesso pode ser realizado não só a partir de computadores pessoais, mas também de dispositivos móveis, o que contribuiu para a popularização do acesso a dados na Web, incluindo vídeo (bate-papo com vídeo, sites como Youtube, por exemplo). Entretanto grande parte dos aparelhos móveis não possui capacidade de processamento e resolução de tela suficientes para exibir conteúdo com alta taxa de dados e em alta definição (vídeo em especial). Uma das áreas que tentam minimizar esses problemas é a adaptação e personalização de conteúdo [1].

Um sistema de adaptação decide a melhor versão de conteúdo para ser apresentada em determinada situação, e a melhor estratégia para se criar essa versão [2]. A personalização, por sua vez, é vista como um caso particular de adaptação, onde os dados são adaptados conforme as preferências ou necessidades de um usuário específico, segundo Barrios et al. [3].

Para que um vídeo seja personalizado de acordo com as preferências de um usuário específico, em geral, é necessário obter informações sobre o mesmo e compará-las com as referidas preferências [3, 4, 5]. Dentre os tipos de informações possíveis de serem obtidas estão as informações semânticas, como por exemplo o assunto que está sendo tratado, os objetos e pessoas presentes na cena, etc. Esse tipo de informção tem atraído a atenção de pesquisadores devido a seu potencial para melhorar tarefas de personalização [5, 6, 7]. Entretanto, a extração de significado semântico é uma tarefa não trivial e muitas vezes subjetiva, visto que cada usuário interpreta o assunto tratado à sua maneira, em geral é necessário que haja interação humana na criação de uma base de dados com informações que possam ser utilizadas por algoritmos que visam realizar a classificação semântica do vídeo de maneira automática [8].

Algumas dessas informações semânticas podem ser obtidas por meio de técnicas de detecção e reconhecimento de faces (identificação, número de pessoas na cena, etc), auxiliando o processo de personalização. Contudo, as técnicas disponíveis para localização de faces em vídeos o fazem realizando análise exaustiva de cada quadro do vídeo, o que é computacionalmente caro [9].

Em um trabalho relacionado, Manzato e Goularte [5] utilizam faces em um sistema de recomendação de vídeos (vide Seção 2). Como o processo de varrer uma base de vídeos à procura de faces relacionadas durante a interação do usuário é computacionalmente caro e demorado, realiza-se um pré-processamento (*off-line*) dos vídeos para extrair quadros que contenham faces, compondo a base utilizada no processo de busca. Contudo, atualmente, esse pré-processamento é manual.

Assim, o objetivo deste trabalho é propor uma técnica que viabilize a extração automática de quadros-chave, de maneira a obter um conjunto de dados menor em comparação com a abordagem de Manzato e Goularte [5], minimizando assim o custo computacional sem comprometer severamente os resultados.

2. TRABALHOS RELACIONADOS

Manzato e Goularte [5] propõem um sistema que utiliza informações semânticas provenientes de faces em quadros de vídeos como um dos subsídios para recomendar vídeos relacionados. Nesse sistema o usuário interage com o vídeo marcando (com tinta eletrônica) uma face de interesse e obtém como reposta uma lista de vídeos que contém aquela face. Durante a busca o sistema compara a face marcada com uma base de faces (quadros de vídeo). Contudo, atualmente, a segmentação de vídeos em quadros contendo faces para compor a base não é um processo automático.

No que se refere apenas a extração de quadros-chave em vídeos digitais, Huang Et al [10] definem quadros-chave como o conjunto de quadros que possuem informação suficiente sobre a cena da qual eles pertencem para possibilitar que, a partir de seus quadros-chave, seja possível construir novamente a cena original, e são importantes pois, no caso ideal, contém todas as

informações relevantes contidas na cena e podem ser processados como imagens estáticas.

Em um vídeo, o conjunto de quadros consecutivos capturados por uma mesma câmera, sem que haja cortes ou mudança abrupta da área que está sendo gravada é chamado de tomada. Alguns autores como Jain et al. [11] e Tang et al. [12] apresentaram um método que utiliza similaridade entre quadros para obtenção de quadros-chave relativos a cada tomada.

Rui et al. [13] propõem uma abordagem temporal, onde se escolhe um quadro-chave a cada intervalo de tempo pré-estabelecido, sem que seja verificado se há ou não mudança de tomada próxima do quadro. Nesse caso a escolha de qual valor será escolhido para o intervalo varia de acordo com o domínio de aplicação.

Como se pode perceber a definição de quadro-chave depende do domínio da aplicação, sendo que a abordagem de Rui et al. [13] é a mais flexível e adequada ao propósito deste trabalho. No contexto específico de obtenção de informação semântica através de detecção facial em vídeos, não foram encontrados trabalhos relacionados, sendo um diferencial desta proposta, a qual também analisa qual o intervalo de tempo ideal para extração dos quadros-chave no referido domínio.

3. DESENVOLVIMENTO

3.1 Extrator de Quadros-Chave

O extrator de quadros-chave foi desenvolvido e implementado em linguagem C, segundo um critério flexível, onde o usuário pode determinar por qual intervalo de tempo ou por qual intervalo de quadros os quadro-chave serão extraídos. Os quadros-chave extraídos são salvos como arquivos JPEG.

Para o processamento dos vídeos, o programa criado utiliza as bibliotecas FFMPEG [14], pois essas permitem exploração das características do vídeo quadro-a-quadro, por exemplo: o tipo do quadro - I, P ou B; instante relativo de tempo de um quadro vídeo. Além disso, a FFMPEG permite acesso aos codecs de vídeo mais comumente utilizados (MPEG, MOV, XVid, etc.) .

O formato JPEG foi escolhido para os quadros-chave por obter boas taxas de compressão para uma grande variedade de imagens e por ser amplamente utilizado. Outra vantagem do JPEG é que quadros do tipo I já são codificados utilizando esse algoritmo, o que torna a extração mais rápida quando se sabe o tipo de quadro que será extraído.

3.2 Detecção Facial

A técnica escolhida para detecção facial nos vídeos foi a proposta por Viola e Jones [9], que realiza a detecção facial utilizando *Haar Classifier* e *Pose Estimator*, por ser uma técnica que apresenta altas taxas de sucesso (de 85 a 95%) e por essa técnica já ter sido implementada em código aberto, na ferramenta Faint [15].

A ferramenta Faint oferece um conjunto de classes Java, que foram utilizadas para a criação de um procedimento para a realização da detecção nos quadros-chave extraídos dos vídeos. Tal procedimento realiza a análise em todos os quadros-chave em busca de faces e armazena as informações das faces detectadas.

As informações armazenadas são a quantidade de faces encontradas, a posição de cada face no quadro-chave e seu tamanho, em pixels. Além disso, também salva uma cópia das faces em formato JPEG para posterior comparação entre as técnicas usadas na extração, com o objetivo de determinar suas respectivas taxas de acerto.

3.3 Testes

Nos testes foram utilizados vídeos da base montada pelo grupo de pesquisa [5, 16], contendo 25 programas telejornais de 25 dias diferentes. Não foram utilizadas bases padronizadas, como TRECVID[1] e MediaMill[2], pois tais bases oferecem seqüências curtas e nem sempre no domínio de telejornais, se opondo às necessidades dos projetos relacionados [5, 16].

O extrator foi aplicado a um conjunto aleatório de 10 vídeos da base segundo os seguintes critérios: um quadro-chave a cada quadro do vídeo; um quadro-chave a cada dez quadros; um quadro-chave a cada dois segundos. Assim, para cada vídeo, foram gerados três sub-conjuntos de quadros-chave.

Figura 1 – Quadro-chave extraído do vídeo

O propósito de executar a extração de um quadro-chave a cada quadro do vídeo foi para criar um modelo com o qual os outros critérios podem ser comparados.

O programa executa a extração dos quadros-chave e gera um arquivo de texto com informações sobre o instante de tempo em que o quadro aparece no vídeo, o posicionamento e tamanho da(s) face(s) encontrada(s), como mostrado nas Figuras 1 e 2.

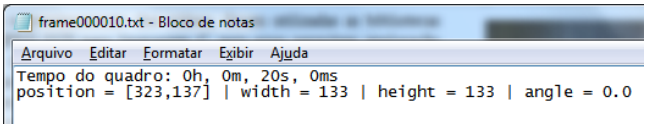

Figura 2 – Arquivo de Texto referente ao quadro-chave

Caso haja mais de uma face detectada no vídeo, são escritas no arquivo as informações de cada face, e no caso de não haver faces, há apenas o tempo do quadro no vídeo. Essas informações podem ser posteriormente utilizadas por outras aplicações como subsídios para tarefas de personalização.

4. RESULTADOS

Nos sub-conjuntos de quadros-chave foi aplicada a detecção facial (via a ferramenta Faint), anotando-se o tempo médio de execução bem como a quantidade média de quadros-chave extraídos, como mostra a Tabela 1.

[1] http://trecvid.nist.gov/
[2] http://www.science.uva.nl/research/mediamill/

Tabela 1 –Tempo médio de execução do programa.

Técnica	Tempo	Quadros
1 Quadro-Chave a cada Quadro	60m36s	4367
1 Quadro-Chave a cada 10 Quadros	8m21s	436
1 Quadro-Chave a cada 2 Segundos	3m4s	89

Após a execução, analisou-se a quantidade de faces encontradas nas 3 técnicas, de modo a determinar o desempenho da utilização de quadros-chave em comparação com a detecção em todos os quadros.

Para essa análise foram considerados apenas os verdadeiros-positivos (faces que existiam e foram encontradas) e falsos-negativos (faces que existiam e não foram encontradas), e ignorados os falsos-positivos (objetos que não eram faces mas o programa interpretou como face) e os verdadeiros-negativos (não-face e não-detectado).

Desse modo é possível comparar a quantidade de faces reais (não outros objetos classificados erroneamente) que foram detectadas segundo cada técnica, e realizar a comparação - apresentada na Tabela 2 - com a média das porcentagens de detecções obtidas.

Tabela 2 – Faces detectadas e perdidas

Técnica	V.P. (%)	F.N. (%)
1 Quadro-Chave a cada Quadro	100	0
1 Quadro-Chave a cada 10 Quadros	98,05	1,95
1 Quadro-Chave a cada 2 Segundos	94,67	5,33

5. CONCLUSÕES

Ao observar a Tabela 1 percebe-se que a média dos tempos de execução do programa cai de modo considerável ao se reduzir a quantidade de quadros analisados, conforme esperado. Desse modo, pode-se comprovar que é de fato computacionalmente mais barato utilizar quadros-chave ao invés de analisar todos os quadros do vídeo, desde que a técnica selecione significativamente menos quadros que o total existente no vídeo.

Para o caso de 1 quadro-chave a cada 2 segundos reduziu-se o número de quadros na ordem de 49 vezes e obteve-se uma redução do tempo na ordem de 20 vezes. Apesar dessa grande redução no número de quadros, a eficiência na detecção de faces foi de 94,67%. Além disso, a Tabela 2 mostra que a taxa de detecções teve queda de 2% ao utilizar intervalos de 10 quadros e de 7% com 2 segundos, o que é relativamente pouco diante da redução do custo computacional demonstrada pela Tabela 1.

Sendo assim, pode-se concluir que a diminuição do número de quadros a serem analisados para detecção facial no domínio de vídeos de telejornais é viável.

Outro benefício deste trabalho é a implementação de uma técnica automática para pré-processamento de vídeos com propósito de extrair quadros que contenham faces, alimentado a base utilizada no processo de busca em aplicações de personalização e recomendação de vídeo – conforme discutido nas seções 1 e 2.

Como trabalhos futuros pretende-se verificar se ocorre queda significativa no desempenho do sistema de recomendação ao se utilizar quadros-chave na detecção facial. Também pretende-se integrar ao sistema de recomendação as informações sobre faces obtidas pela ferramenta.

AGRADECIMENTOS

Os autores agradecem à FAPESP pelo apoio financeiro a este projeto de Iniciação Científica.

REFERÊNCIAS

[1] Lu, Y. H., Ebert, D. S., Delp, E. J. 2006. *Resource-Driven Content Adaptation*. School of Electrical and Computer Engineering Purdue University West Lafayette, Indiana.

[2] Lum, W. Y., Lau, F. C. M. A. 2002. *Context-Aware Decision Engine for Context Adaptation*. IEEE Pervasive Computing, v. 1, n. 3, 41–49p.

[3] Barrios, V. M. G., Mödritscher, F., Gütl, C. 2005. *Personalization versus Adaptation? A User-centred Model Approach and its Application*. Proceedings of I-KNOW'05. 120-127p.

[4] Bertini, M. Bimbo, A. D., Cucchiara, R., Prati, A. 2004 *Semantic Video Adaptation based on Automatic Annotation of Sport Videos*. Proceedings of the 6th ACM SIGMM. 291-298p.

[5] Manzato, M. G., Goularte, R. 2009. *Supporting multimedia recommender systems with peer-level annotations*. Brazilian Symposium on Multimedia and the Web. 202-209p.

[6] Ulges, A., Koch, M., Schulze, C., Breuel T. 2008. *Learning TRECVID'08 high-level features from YouTube#8482*. Proc. of TRECVID 2008.

[7] Xinbo G., Yimin Y., Dacheng T., Xuelong L. 2009. *Discriminative optical flow tensor for video semantic analysis,* Computer Vision and Image Understanding, v.113 n.3, 372-383p.

[8] Jiang H, Elmagarmid A.K. 2002. *Knowledge and Data Engineering*. IEEE Transactions on. Dept of Comput. Sci., Purdue Univ., West Lafayette, Indiana.

[9] Viola, P., Jones, M.J. 2004. *Robust real-time face detection*. International Journal of Computer Vision. 137-154p.

[10] Huang, K.-S., Chang, C.-F., Hsu, Y.-Y., Yang, S.-N. 2005. *Key Probe: a technique for animation keyframe extraction*. The Visual Computer 21 8-10, 532-541p.

[11] Jain, A., Vailaya, A., Xiong, W. 1998. *Query by Video Clip*. Proceedings of International conference on Pattern Recognition. 909-911p.

[12] Tang, S., Zhang, Y.-D., Li, J.-T., Pan, X.-F., Xia, T., Li, M., Liu, A., Bao, L., Liu1, S.-C., Yan, Q.-F., Tan, L. 2006. *TRECVID 2006 Rushes Exploitation By CAS MCG*. Chinese Academy of Sciences, Beijing.

[13] Rui, Y., Huang, T.S., Mehrotra, S. 1998. *Exploring video structure beyond the shots*. Proceedings of IEEE International Conference on Multimedia Computing and Systems (ICMCS), Texas, USA, 237–240p.

[14] http://ffmpeg.org – FFMPEG, acessado em 25/04/2010.

[15] http://faint.sourceforge.net/ - Faint – The Face Annotation Interface, acessado em 25/04/2010.

[16] Coimbra, D. B., Goularte, R. 2009. *Identificação de Cenas em Vídeos Digitais Utilizando Características Audiovisuais*. Brazilian Symposium on Multimedia and the Web. v. 2. 43-46p.