Interação Humano-Computador Utilizando Técnicas de Visão Computacional

Alessio Delazari[1], Jonas Rotta[1], Antonio Carlos Sobieranski[1], Eros Comunello[1]

[1] LAPIX - Laboratório de Processamento de Imagens e Computação Gráfica
- The Cyclops Group
INE - Departamento de Informática e Estatística
Universidade Federal de Santa Catarina - UFSC - Brasil

{alessio, jonas, asobieranski, eros}@cyclops.ufsc.br

Abstract. *In this paper a novel computational approach for human-machine interaction using computer vision techniques is presented. In order to propose a low-cost methodology, only common web-cams are used, where their frames are captured and processed using computer vision techniques. Two application contexts are presented: the environment control using actions from the user defined by free-hand; targets identification to apply the user to the environment - extended reallity. The preliminary results we obtained have shown the viability of the proposed approach for the presented contexts.*

Resumo. *Neste artigo será apresentado uma metodologia computacional para a interação humano-computador utilizando técnicas de visão computacional. Para tal, serão utilizadas imagens providas a partir de uma webcam juntamente com técnicas de processamento digital de imagens, com o objetivo de ser uma metodologia de baixo custo. Dois contextos serão explorados: controle do ambiente utilizando ações do usuário à mão livre; identificação de objetos de interesse para imersão do usuário no ambiente - realidade estendida. Os resultados preliminares demonstram a viabilidade da metodologia proposta para os contextos apresentados.*

1. Introdução

A área de interação humano-computador vem se destacando nos últimos anos como uma tecnologia promissora. Seu maior impulso deve-se principalmente a área de jogos eletrônicos, que cada vez mais exigem a inserção do jogador em ambientes de realidade virtual. Entretanto, muitos outros ambientes vêm-se utilizando desta tecnologia também para aplicações de vídeo e tele-conferência, auxílio portadores de deficiência física, e outras aplicações que objetivem a simples utilização do mouse e teclado.

Dentre as diversas técnicas e formas de interação, entretanto, destacam-se os DataGlovers, um equipamento composto por uma luva com sensores para a imersão em realidade virtual, sendo que com tais sensores é possível a aquisição de informações espaciais e de flexionamento de cada um dos dedos [Sturman and Zeltzer 1994]. Outros dispositivos são os óculos para visualização em perspectiva 3D, que basicamente podem ser divididos em 3 categorias: anáglifo, utilizando-se de filtros de cores complementares entre

vermelho e azul, onde a imagem visualizada proporciona sensações de 3D e imersão espacial; polarizados, onde as imagens são separadas através da luz e as suas características de cores não são alteradas, mesmo que com uma pequena perda de luminosidade; alternativos, as imagens da esquerda e da direita são apresentadas seqüencialmente, e sincronizadas através de óculos dotados de obturadores de cristal líquido, de modo que cada olho perceba a sua imagem correspondente. Algumas destas tecnologias são atualmente empregadas em monitores convencionais LCD, TV e cinemas 3D de última geração, e devido a sua elevada freqüência de operacionalização são imperceptíveis ao olho humano, proporcionando a imersão em um ambiente 3D [Terra 2003]. Outros meios de interação humano-computador, assim como mouse e teclados, são totalmente físicos, e o movimento são matematicamente traduzidos em uma seqüência binária. Este tipo de interação é diretamente traduzida para um ambiente virtual, sendo algo naturalmente realizado pelas gerações mais novas através das gerações medianas de videogames.

As tecnologias anteriormente especificadas, de certa forma as mais recentes, envolvem um custo elevado para a sua aquisição. Uma nova modalidade de interação humano-computador que vem emergindo é através da combinação de recursos já existentes em qualquer computador residencial: as webcams. Webcams capturam continuamente as imagens contidas no ambiente, e através de sua seqüencia temporal, são codificadas em vídeo. Através da utilização de técnicas de processamento digital de imagens (PDI), a identificação de ações pré-definidas dos usuários podem ser realizadas através da codificação de movimentos em ações. Alguns trabalhos correlatos podem ser encontrados na literatura: em [Bandera et al. 2009] foi proposto o desenvolvimento de uma interface de interação homem-robô, com uma abordagem a dois níveis para reconhecer gestos que são compostos de trajetórias seguidas de diferentes partes do corpo. Em um primeiro nível, trajetórias individuais são descritas por um conjunto de pontos-chave. Em um segundo nível, os gestos são caracterizados através de propriedades globais das trajetórias que as compõem. Em [Rodriguez et al. 2004] foi apresentado uma nova abordagem para a medição da similaridade entre curvas 3D. Este trabalho permite a possibilidade de usar strings, onde cada elemento é um vetor em vez de apenas um símbolo. Foram propostas duas abordagens diferentes para a representação de curvas 3D. Uma possibilidade é a de representar uma curva em 3D como duas curvas 2D, sendo uma projeção da curva 3D no plano xy, e outro no plano yz. Para os casos invariância e rotação geométrica utilizou-se uma segunda abordagem para a representação simbólica da curva 3D usando a curvatura e a tensão como a sua representação simbólica. Em [Chen et al. 2005] faz uma consideração importante na similaridade baseada em recuperação de trajetórias de objetos em movimento é a definição de uma função de distância. As funções distância existentes são geralmente sensíveis ao ruído, desvios e dimensionamento de dados que geralmente ocorrem devido a falhas de sensores, erros de técnicas de detecção, os sinais de perturbação, e taxas de amostragem diferentes. Limpeza de dados para eliminar estes nem sempre é possível. Neste trabalho uma nova métrica de distância foi apresentada, chamada Edit Distance Real (EDR), que é robusto contra estas imperfeições de dados. Análise e comparação dos EDR com outras funções distância populares, tais como a distância Euclidiana, Dynamic Time Warping (DTW), Edit distância com o Real Penalty (ERP), e Longest comum subseqüências (LCSS), indicam que EDR é mais robusta do que a distância euclidiana, DTW e ERP, e é em média 50% mais precisos que LCSS. Também desenvolvemos três técnicas de poda para melhorar a eficiência de recuperação de EDR e

mostrar que essas técnicas podem ser combinadas de forma eficaz em uma pesquisa, aumentando o poder de poda de forma significativa. Os resultados experimentais confirmam a maior eficiência dos métodos combinados.

Neste artigo, uma nova abordagem para a identificação de objetos de interesse e ações do usuário são propostas. Primeiramente, um framework para processamento de imagens adequado para o processamento de vídeo é proposto. Este framework possibilita a identificação de objetos de interesse nas seqüências de vídeo, e computar as suas trajetórias ao longo do tempo. Acima deste framework, uma nova camada para a aplicação das seqüências de objetos identificados nas imagens em ações foi desenvolvido. Esta camada possibilita a interação do usuário com o computador pela simples utilização da webcam em ações.

O presente artigo está organizado da seguinte forma: seção 2 descreve a metodologia proposta e as técnicas de processamento digital de imagens utilizadas. Seção 3 descreve os resultados preliminares obtidos para 2 contextos de aplicação: controle de ações e imersão do usuário no ambiente. Por fim, na seção 4 serão apresentados as conclusões, discussões e trabalhos futuros pretendidos para a metodologia proposta.

2. Desenvolvimento

No diagrama da Figura 1 é apresentado uma visão geral da metodologia proposta para a utilização no contexto de interação humano-computador. As etapas iniciais da aquisição dos frames do dispositivo e o processamento prévio é realizado por:.

- Uma webcam comum pode ser utilizada para a metodologia proposta. Para a captura dos frames foi utilizada a biblioteca OpenCV, que já possui uma camada de abstração implementada que possibilita a aquisição dos frames da maioria das câmeras comerciais existentes. Outra vantagem é a portabilidade, sendo que esta plataforma foi desenvolvida em Linux utilizando a linguagem C++, sendo passiva de ser compilada para a plataforma Windows ou outro sistema operacional;
- Através da captura dos frames da imagem, técnicas de processamento digital de imagens são aplicadas para a identificação dos targets nas imagens. Por targets, a abordagem proposta generaliza, uma vez que pode ser um objeto de interesse na cena (como uma bola vermelha, uma caneta, etc), ou qualquer movimentação que é executada pelo usuário;

A identificação dos targets de interesse pode ser realizada até então de duas formas pela metodologia proposta. Quando nenhum target de interesse é definido, os targets podem ser qualquer alteração que ocorrer no ambiente. Isso pode ser identificado através de sucessivas subtrações realizadas entre as cenas com o objetivo de identificar tais alterações. Para a segunda abordagem, um target de interesse é definido através de suas características de cores, e procurados por padrões similares em cada cena subseqüente. Isso é realizado através da transformada da distância Euclidiana $||u - \mu||$ entre os padrões de cores do objeto de interesse e a imagem.

Uma vez que os targets de interesse estejam identificados, um processo de melhoria na informação é realizado. Primeiramente, um algoritmo de dilatação é executado sobre a imagem binária correspondendo aos objetos da cena. Este algoritmo tem como objetivo a correção de pequenas falhas e ruídos na imagem. Um algoritmo de componentes conexos (popularmente conhecido por *labeling*) efetua uma varredura em cada

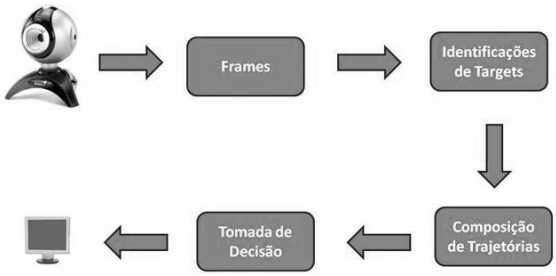

Figure 1. Diagrama geral da metodologia proposta

frame do vídeo e extraindo os meta-dados dos frames. Outras características também são utilizadas, tais como o índice do objeto na cena (int) e a coordenada central do objeto correspondendo ao seu centro de massa (coordenada x e y).

Uma vez que os meta-dados estão extraídos, os mesmo são armazenados em uma estrutura FIFO de tamanho pré-definido (default 10). Através da análise individual de cada objeto e pela procura deste objeto das cenas anteriores, ocorre a composição das trajetórias deste objeto.

As trajetórias são analisadas individualmente, através desta análise ações pré-definidas podem ser aplicadas sobre a cena. Para a metodologia proposta, suas abordagens são exploradas:

- Controle do Ambiente: um objeto que ao longo do tempo possui suas coordenadas x e y alteradas, podem ser codificado na translação de uma janela ou um menu pré-definido. Dois objetos O_1 e O_2, que mantém entre si a distância $D = ((O_1.x - O_2.x) + (O_1.y - O_2.y))^{\frac{1}{2}}$, e essa distância ao longo do tempo aumenta e ambos os objetos diminuem em y, essa ação pode ser codificada em realizar um zoom in no vídeo que está sendo amostrado. A recíproca também pode ser realizada, quando a distância reduz no tempo e ambos os objetos aumentam em y, caracterizando assim uma ação zoom out.
- Um ou mais objetos, que são identificados na cena, podem ser "ligados" por uma linha virtual. Conforme os objetos se movem ou se afastam na cena, a linha acompanha esses objetos através de suas coordenadas de centro de massa x e y.

3. Resultados

Para as duas abordagens exploradas nestes trabalhos, resultados preliminares interessantes foram obtidos. Na figura 2 ambas as abordagens são demonstradas. Na figura à direita, o usuário está efetuando ações sem a utilização de qualquer objeto, somente a mão livre. Este é o ambiente utilizado para controlar a janela de vídeo, tais como ações de translação e zoom in/out do vídeo que está rodando.

Na figura à esquerda, uma representação de ambiente imersivo é representada. Ao identificar 2 objetos de interesse, o protótipo desenha uma linha virtual que liga esses 2 objetos, e um terceiro objeto caracterizado por uma bola cai sobre a linha. O objetivo do usuário é equilibrar a bola em cima da linha como um jogo de equilíbrio. Dentre os 2 objetos, a coordenada y que estiver abaixo uma da outra, solicita para a bola deslocar-se

Figure 2. Resultados da metodologia proposta. A direita, identificação de movimentos a mão livre, à esquerda, game do equilíbrio utilizando 2 targets

sobre a linha, no sentido esquerdo ou direito. Ocultando os targets de interesse, o game reinicia.

4. Conclusão

O presente artigo apresentou os resultados preliminares da aplicação combinada de recursos tecnológicos de baixo custo, como uma simples webcam, e a técnicas de processamento digital de imagens. Com isso, foi possível identificar objetos de interesse e mesmo movimentações do usuário utilizando mão livre. Os resultados obtidos demonstram-se promissores para a utilização em larga escala em ambientes imersivos ou realidade estendida.

Como trabalhos futuros pretende-se a extensão da metodologia proposta para a identificação em outros contextos de aplicação, como informação facial, considerada de complexa identificação. Outras melhorias devem ser efetuadas na questão das trajetórias, pois o algoritmo de tracking é somente baseado na ligação das coordenadas centrais de cada objeto de interesse.

References

Bandera, J. P., Marfil, R., Bandera, A., Rodríguez, J. A., Molina-Tanco, L., and Sandoval, F. (2009). Fast gesture recognition based on a two-level representation. *Pattern Recogn. Lett.*, 30(13):1181–1189.

Chen, L., Özsu, M. T., and Oria, V. (2005). Robust and fast similarity search for moving object trajectories. In *SIGMOD '05: Proceedings of the 2005 ACM SIGMOD international conference on Management of data*, pages 491–502, New York, NY, USA. ACM.

Rodriguez, W., Last, M., Kandel, A., and Bunke, H. (2004). 3-dimensional curve similarity using string matching. *Robotics and Autonomous Systems*, 49(3-4):165–172.

Sturman, D. J. and Zeltzer, D. (1994). A survey of glove-based input. *IEEE Comput. Graph. Appl.*, 14(1):30–39.

Terra (2003). Como funcionam os óculos 3D usados no cinema. http://noticias.terra.com.br/ciencia/interna/0,,OI233515-EI1426, 00.html.