9º FEPEG FÓRUM DE ENSINO, PESQUISA, EXTENSÃO E GESTÃO
23 A 26 SETEMBRO DE 2015
Campus Universitário Professor Darcy Ribeiro
TRABALHOS CIENTÍFICOS APRESENTAÇÕES ARTÍSTICAS E CULTURAIS DEBATES MINICURSOS E PALESTRAS
ISSN 1806-549X
A HUMANIZAÇÃO NA CIÊNCIA, TECNOLOGIA E INOVAÇÃO

Modelagem Gráfica para Reconhecimento de Ações Humanas em Sequências de Nuvens de Pontos Usando Invariantes Geométricos

Gustavo Willer Ferreira

Introdução

Com a popularização dos sensores de profundidade e extração de esqueletos, na forma de grafo, em nuvens de pontos, diversos dispositivos são adaptados para jogos cada vez mais realistas onde o jogador usa os próprios membros para interagir com o ambiente virtual. Exemplo de tais dispositivos é o sensor Kinect, que acompanha o MS-Xbox [1]. Além da aplicação em games, a comunidade científica de visão computacional tem explorado o amplo campo de aplicações do reconhecimento de ações humanas, como em vigilância, onde gestos são analisados para detectar acidentes domésticos com crianças e idosos ou gestos suspeitos de potencial violência [2]. Outras aplicações incluem a análise de desempenho de atletas e desenvolvimento de dispositivos para atuação em realidade virtual, com aplicação em ambientes industriais e automação de utensílios domésticos.

O primeiro obstáculo do reconhecimento de ações se manifesta pelas diferenças culturais entre as pessoas, esse fato é percebido na variação da velocidade de execução de uma ação ou nas variadas sequências de poses que uma mesma ação pode determinar. O presente trabalho tem como objetivo estudar os algoritmos de reconhecimento de ações, conhecer as principais bases de dados públicas produzidas com esqueletos extraídos de dispositivos de imagem de profundidade de tempo real, bem como conhecer e contribuir no sentido de melhorar as atuais taxas de reconhecimento de gestos.

Material e Métodos

As ações são compostas por uma sequência de poses conforme ilustrado na Figura 1, em que geralmente a pose final coincide com a inicial. O número de possibilidades de poses distintas muito grande, contudo é possível identificar poses predominantes, que aparecem em diversas ações, chamadas *poses salientes*. Cada ação será, então, expressa por sequências dessas *poses salientes*. Para esse processamento do reconhecimento das *poses salientes*, temos que alcançar três objetivos principais: A descrição da pose, a identificação da pose, e o reconhecimento da sequência de poses como um gesto, ou ação [3].

A descrição da pose, em nosso trabalho, é feita através de matrizes de distâncias que provêm a invariância no tempo e uma maior robustez, de forma que poses equivalentes são mapeadas como o mesmo ponto referente à matriz de distância [4]. A identificação da pose é realizada através da construção de um modelo de observação para cada pose na fase de treinamento. E nessa fase também é construído um modelo de transição referente a cada classe de ação para a identificação dos gestos. Enquanto que o modelo de observação é construído como uma distribuição gaussiana para cada pose saliente, o modelo de transição é construído como uma cadeia escondida de Markov.

O método proposto consta de uma fase de treinamento, onde um conjunto de ações previamente conhecidas é usado para construir os modelos de observação e de transição. De posse desses modelos, uma fase de testes usa um conjunto de sequências sem o conhecimento prévio das ações a que se referem com o objetivo de associar automaticamente cada sequência ao nome da ação correta que ela descreve. As etapas da seção de treinamento constam de descrição da pose, redução da dimensão, agrupamento em *poses salientes*, e modelagem gráfica da dinâmica das ações. A etapa de testes consta da decodificação de ações.

A. Descrição da pose: Uma esqueleto S obtido com o dispositivo kinect, que tem 20 juntas representando a postura de um sujeito, é mapeado para uma matriz 20x20 onde cada elemento (i,j) armazena a distância entre as juntas pi e pj, descrevendo então 400 valores de distância para representar uma única pose. A vantagem das matrizes de distância é que duas ações equivalentes terão as mesmas representações, mesmo se capturadas em orientação distinta pelo dispositivo [4].

B. Redução de dimensão: As poses são inicialmente descritas num espaço vetorial de dimensão n=400. Isso requer um maior custo computacional para todo o processamento. Então, usando PCA (análise de componentes principais), é feita uma redução da dimensão dos vetores descritores das poses. Essa análise PCA permite identificar redundâncias na descrição pela identificação de correlação entre os dados. Havendo correlação significa que os dados podem ser descritos, sem perda, em uma dimensão menor.

C. Agrupamento: O aprendizado das *poses salientes* é feito através da técnica de clusterização Kmeans, que consiste em dividir o conjunto em k grupos contendo pontos próximos entre si. Este agrupamento usa uma técnica de maximização da expectativa onde um particionamento inicial é escolhido randomicamente e um processo iterativo converge para o particionamento ideal segundo a condição inicial. A implementação usual do Kmeans, como no Matlab, faz uma escolha inicial randômica do particionamento inicial, de maneira que a repetibilidade do método fica comprometida nos testes. Para garantir a repetibilidade e ter consistência nos testes, foi desenvolvido uma versão do Kmeans, onde o particionamento inicial é feito de forma determinística.

D. Modelagem gráfica: A representação gráfica da dinâmica das ações é construída nos modelos de observação e de transição. O modelo de observação é obtido a partir de cada agrupamento, ou *cluster*, como sendo uma distribuição gaussiana que melhor explique a distribuição dos dados no cluster. O modelo de transição é construído como um grafo, onde cada nó representa uma *pose saliente* e as arestas representam transições entre *poses salientes*. Um caminho no grafo é então interpretado como uma sequência de poses, descrevendo uma ação.

E. Decodificação de ações: Para o reconhecimento de ações na etapa de testes, sequências desconhecidas são submetidas ao modelo gráfico obtido no treinamento. Essa decodificação é implementada com uma técnica de programação dinâmica chamada *Virtebi.* Bases de dados públicas são usadas para verificar a precisão do reconhecimento nessa fase e fazer a validação do método.

Resultados e Discussão

Usamos uma base de dados pública com 567 sequências capturadas com o sensor Kinect. Os dados constam de 20 diferentes ações, executadas por 10 sujeitos distintos, sendo que cada sujeito executou cada ação por três vezes. Na etapa de treino dividimos em 3 grupos de treinamento com todas as sequências de poses, sendo que em cada grupo retiramos um dos 3 exemplos executados por cada sujeito. Na etapa de testes, para cada grupo de treinamento, usamos as sequências de poses que estiveram fora do treino e executamos a decodificação das ações. A tabela 1 apresenta os resultados obtidos considerando diversos parâmetros para a dimensão dos descritores e quantidade de poses salientes. O melhor resultado obtido com ajuste dos parâmetros é apresentado em detalhes no formato de uma matriz de confusão, conforme a Figura 2.

Considerações finais

Nossos resultados preliminares mostram que a técnica utilizada é consistente e permite uma taxa de reconhecimento que se aproxima dos melhores resultados apresentados na literatura. A continuação do trabalho visa melhorar as taxas de reconhecimento com aprimoramento na construção dos modelos de observação e transição. O modelo de observação ainda apresenta algumas inconsistências, especialmente com a degeneração da matriz de covariância para alguns clusters e acreditamos que a correção dessa degeneração permitirá maiores taxas de reconhecimento. Além disso, no modelo de transição, testes com ajustes de parâmetros serão feitos com objetivo de otimizar os resultados.

Agradecimentos

FAPEMIG e Pró-Reitoria de Pesquisa da Unimontes pela concessão das bolsas de iniciação científica.

Referências

[1] Li, W., Zhang, Z., & Liu, Z. (2010, June). Action recognition based on a bag of 3d points. In Computer Vision and Pattern Recognition Workshops (CVPRW), 2010 IEEE Computer Society Conference on (pp. 9-14). IEEE.

[2] Cao, L., Liu, Z., & Huang, T. S. (2010, June). Cross-dataset action detection. In Computer vision and pattern recognition (CVPR), 2010 IEEE conference on (pp. 1998-2005). IEEE.

[3] Miranda, L., Vieira, T., Martinez, D., Lewiner, T., Vieira, A. W., & Campos, M. F. (2012, August). Real-time gesture recognition from depth data through key poses learning and decision forests. In Graphics, Patterns and Images (SIBGRAPI), 2012 25th SIBGRAPI Conference on (pp. 268-275). IEEE.

[4] Vieira, A. W., Lewiner, T., Schwartz, W. R., & Campos, M. (2012, November). Distance matrices as invariant features for classifying MoCap data. In Pattern Recognition (ICPR), 2012 21st International Conference on(pp. 2934-2937). IEEE.

ISSN 1806-549X

Tabela 1. Taxa de reconhecimento das ações, sendo as vinte classes de ações executadas por três diferentes exemplos de cada sujeito. A coluna representa o numero de dimensões e as linhas o número de *poses salientes* utilizadas.

Poses salientes	**Número de dimensões**									
	2	**3**	**4**	**5**	**6**	**7**	**8**	**9**	**10**	**11**
210	57,70	70,87	79,04	84,23	84.42	82,93	79,78	80,33	80,89	63,64
215	60,67	74.40	79,22	83,67	85,34	84,42	83,36	82,93	77,92	53,99
226	57.51	71,80	77,55	84,42	82,93	**86,09**	82,56	83,86	57,51	46,20

*Os valores estão expressos em porcentagem (%).

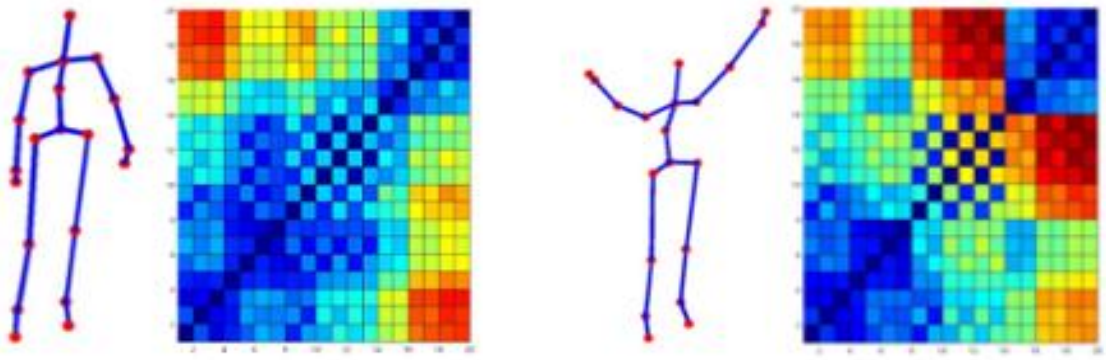

Figura 1. Exemplos de modelo gráfico do esqueleto e matriz de distância. Dois esqueletos obtidos com o Kinect são apresentados, tendo ao lado a respectiva matriz de distancia ilustrada com os valores numa escala de intensidade, onde juntas mais próximas são representadas em azul e juntas mais distantes apresentadas em vermelho.

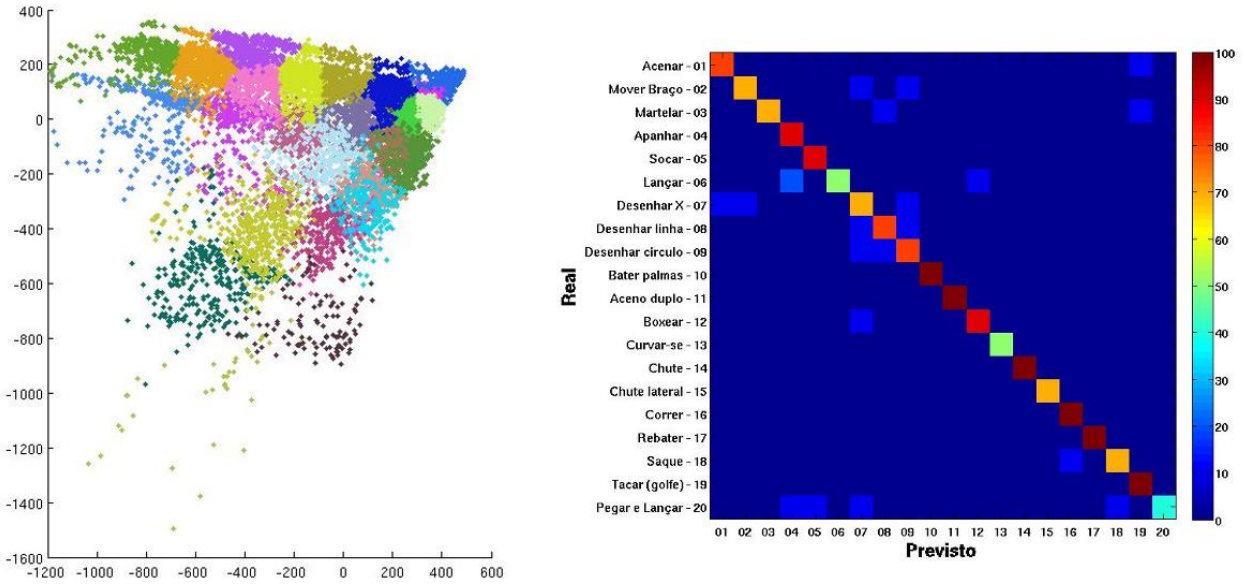

Figura 2. A imagem da esquerda mostra um exemplo do conjunto de descritores projetados no espaço de componentes principais. A imagem da direita mostra uma matriz de confusão com detalhamento de nossos resultados preliminares por cada tipo de ação da base de dados, onde a diagonal principal representa os acertos do algoritmo por ação conforme escala de 0 a 100% representada numa escala de cores de azul a vermelho.