Human-Computer Interaction Based on Visual Recognition using Volumegrams of Local Binary Patterns

Ana I. Maqueda, Carlos R. del-Blanco, Fernando Jaureguizar, Narciso García

***Abstract*—A robust hand-gesture recognition system based on a novel descriptor called Volumetric Spatiogram of Local Binary Patterns is presented, which allows a more natural input interface for simulating a mouse. The recognition stage based on Support Vector Machines triggers different mouse functions depending on the recognized gesture.**

I. Introduction

In recent years, hand gesture recognition systems based on vision have undergone a increasingly popularity due to their wide range of potential applications in the field of Human-Computer Interaction (HCI), such as multimedia application control [1], video-games [2], and medical systems. These interfaces are considered more friendly, natural, and intuitive for the user than traditional HCI devices (mouse, keyboard, etc.). On the other hand, the fact that most of the consumer devices are supplied with color cameras has also motivated the growth of HCI systems based on hand gesture recognition.

In order to recognize dynamic gestures, some works model the temporal information through spatio-temporal features, which are used along with a general classifier or a template-matcher [3]. The extracted features should be invariant to illumination, scale changes, rotations and translations. Image descriptors are used to represent the 2D visual content of image regions [4], which are suitable for static gestures. However, dynamic gestures must take into account the temporal dimension, which increases the complexity. For this purpose, some video descriptors have been developed [5], which tend to excessively simplify the information of the video regions.

We propose a robust vision-based hand gesture recognition system that provides a more natural interface to imitate mouse-like pointing devices. The key contribution of the system is a novel and highly discriminative video descriptor, called Volumetric Spatiogram of Local Binary Patterns (VS-LBP). This descriptor is not only computationally efficient and robust to dramatic illumination changes, but also provides much richer spatio-temporal information than other existing approaches. The VS-LBP descriptors are used as input of a set of Support Vector Machine classifiers (SVM) to perform the recognition of a predefined set of hand-gestures.

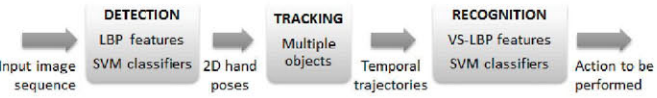

Fig. 1. Proposed hand-gesture recognition system.

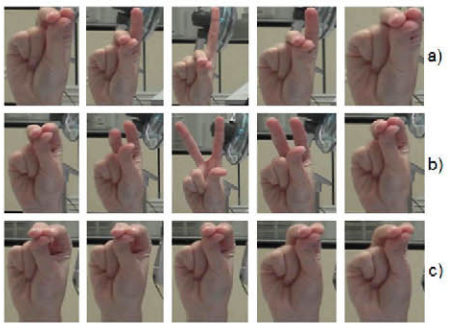

Fig. 2. Hand-gesture samples. (a) Left-click. (b) Right-click. (c) Cursor.

II. System description

The proposed hand-gesture recognition system can be decomposed into three phases: detection, tracking and recognition, as shown in Fig. 1. The detection phase uses a set of SVM classifiers to detect determined static hand poses. The classifiers use Local Binary Patterns (LBP) [6] as input features to perform a fast and efficient detection, and adopts a multi-scale sliding window approach to be robust to scale changes. These detections are used as input of a multiple object tracker [7], which is robust to erroneous, distorted and missing detections, to generate a trajectory of hand poses. The recognition phase uses the tracked hand regions to compute VS-LBP descriptors, which have a much richer information than the LBP descriptor, discriminating the different dynamic hand gestures accurately. The VS-LBP descriptors are delivered to a bank of SVMs that performs the gesture recognition. Depending on the recognized gesture, (see Fig.2) the system executes different actions (cursor motion, left-click, and right-click) that imitates the functions of a mouse device.

III. Feature extraction

The VS-LBP descriptor is a mayor extension of the LBP descriptor that includes global spatial information to be more

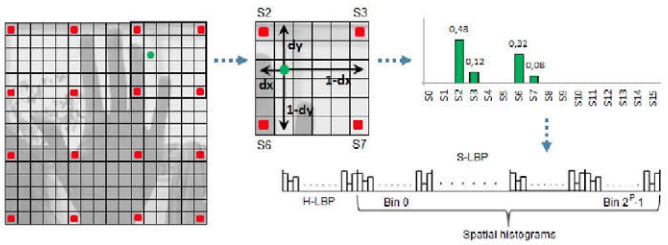

Fig. 3. Each one of the coordinates of an H-LBP bin contributes with different weights (bilinear interpolation) to a spatial histogram. The spatial histograms related to the H-LBP bins and the H-LBP itself are concatenated to form the S-LBP descriptor.

discriminative, and temporal information to deal with dynamic gestures. The algorithm to compute the VS-LBP can be divided into three steps. The first step selects a 2D spatial region obtained in the tracking phase, and then computes the LBP descriptor to obtain a histogram of Local Binary Patterns H-LBP. The second step computes a histogram of spatial coordinates per each bin of the H-LBP. This is carried out by extracting the coordinates of all the LBP patterns that have contributed to a specific H-LBP bin. These coordinates are used to compute a histogram whose range of spatial coordinates is quantized to shorten its length and keep the computational cost manageable. A bilinear interpolation approach is used to compute the contribution of every coordinate to the quantized histogram, which increases the robustness against slight image translations and the image grid effect. The H-LBP itself and the set of spatial histograms are all concatenated to form a super-descriptor called Spatiogram of Local Binary Patterns (S-LBP) (see Fig. 3).

The last step consists of adding temporal information to the S-LBP framework by sliding a spatio-temporal (volumetric) window to analyze the video regions. A sub-sampling scheme is applied in each spatio-temporal window to reduce the number of 2D spatial regions, which also allows to deal with variations in the execution speed of the hand gestures by considering several sub-sampling steps. Next, an S-LBP descriptor is computed for each temporally-sampled 2D spatial region. Lastly, all the computed S-LBP descriptors are concatenated to form the final descriptor, called Volumetric Spatiogram of Local Binary Patterns. This descriptor includes spatio-temporal information in an efficient and compact way that makes it highly discriminative for dynamic hand gestures.

IV. Results

The proposed system has been tested in a database (https://sites.google.com/site/visualgestrecog/) composed by 6 long video sequences, in which different people perform different dynamic hand gestures per sequence: cursor motion, left-click, and right-click gestures, in addition to other non-representative gestures and transitions. There are also two static hand gestures, open palm and fist, which are used to activate and deactivate the dynamic hand gesture recognition. Our algorithm has been compared to two well-known video descriptors also based on LBP descriptor, Volume Local Binary Patterns (VLBP) and Local Binary Patterns from Three Orthogonal Planes (LBP-TOP) [8]. Table I shows the accuracy results (accuracy(%) $= \frac{\text{Total number of correct gestures}}{\text{Total number of gestures}} \times 100$) for each video sequence and for each video descriptor. Observe that the proposed algorithm outperforms the other approaches, achieving the best accuracy scores in all the sequences.

TABLE I
Accuracy results

Sequence	Accuracy (%)		
	VS-LBP	VLBP	LBP-TOP
seq1	87.3	78.8	77.7
seq2	85.4	80.1	79.1
seq3	86.8	78.4	77.9
seq4	83.2	78	76.7
seq5	89.2	81.2	79.5
seq6	81.6	77.6	77.1

V. Conclusion

A natural and reliable human-computer interface based on a visual hand-gesture recognition approach that simulates a mouse-like pointing device is presented in this paper. The key element of the system is a novel and highly discriminative spatio-temporal descriptor called Volumetric Spatiogram of Local Binary Patterns, which is used to feed a bank of Support Vector Machine classifiers that recognize a set of hand-gestures. The recognized gestures trigger different mouse functions that allows the control of multimedia devices.

References

[1] S.-H. Lee, M.-K. Sohn, D.-J. Kim, B. Kim, and H. Kim, "Smart tv interaction system using face and hand gesture recognition," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2013, pp. 173–174.

[2] M. Pourazad, A. Bashashati, and P. Nasiopoulos, "A random forests-based approach for estimating depth of human body gestures using a single video camera," in *IEEE International Conference on Consumer Electronics (ICCE)*, Jan 2011, pp. 649–650.

[3] M. Abid, L. Melo, and E. Petriu, "Dynamic sign language and voice recognition for smart home interactive application," in *IEEE International Symposium on Medical Measurements and Applications Proceedings (MeMeA)*, May 2013, pp. 139–144.

[4] M. Ozdamar and R. Edizkan, "Evaluation of image descriptors in subspace-based classifiers for traffic sign recognition," in *Signal Processing and Communications Applications Conference (SIU)*, Apr 2014, pp. 574–577.

[5] S. Umakanthan, S. Denman, S. Sridharan, C. Fookes, and T. Wark, "Spatio temporal feature evaluation for action recognition," in *International Conference on Digital Image Computing Techniques and Applications (DICTA)*, Dec 2012, pp. 1–8.

[6] T. Ojala, M. Pietikainen, and T. Maenpaa, "Multiresolution gray-scale and rotation invariant texture classification with local binary patterns," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 7, pp. 971–987, Jul 2002.

[7] C. del Blanco, F. Jaureguizar, and N. Garcia, "An efficient multiple object detection and tracking framework for automatic counting and video surveillance applications," *IEEE Transactions on Consumer Electronics*, vol. 58, no. 3, pp. 857–862, Aug 2012.

[8] G. Zhao and M. Pietikainen, "Dynamic texture recognition using local binary patterns with an application to facial expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 29, no. 6, pp. 915–928, June 2007.