ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

FINGeR: Framework for Interactive Neural-based Gesture Recognition

German Ignacio Parisi, Pablo Barros, and Stefan Wermter *

University of Hamburg - Department of Computer Science
Vogt-Koelln-Strasse 30, D-22527 Hamburg - Germany
http://www.informatik.uni-hamburg.de/WTM/

Abstract. For operating in real world scenarios, the recognition of human gestures must be adaptive, robust and fast. Despite the prominent use of Kinect-like range sensors for demanding visual tasks involving motion, it still remains unclear how to process depth information for efficiently extrapolating the dynamics of hand gestures. We propose a learning framework based on neural evidence for processing visual information. We first segment and extract spatiotemporal hand properties from RGB-D videos. Shape and motion features are then processed by two parallel streams of hierarchical self-organizing maps and subsequently combined for a more robust representation. We provide experimental results to show how multi-cue integration increases recognition rates over a single-cue approach.

1 Introduction

The recent interest in low-cost range sensors using structured light technologies[1] has led to promising results for the recognition of hand gestures [1]. The combination of color and depth information (RGB-D) has been shown to increase robustness under varying light conditions and reduce computational effort [2]. However, the human brain continues to outperform vision-based applications in terms of performance and accuracy. This suggests that the introduction of principles related to the dynamics of the neural visual system can represent a good strategy for computational implementations aiming to address visual tasks that a human observer can achieve effortlessly [3]. While the use of biologically inspired neural architectures to learn human gestures and actions was proposed by a number of approaches [4][5][6], the processing of different visual cues for learning representations of gestures using hierarchical models has not yet been explored extensively.

We present an interactive framework for learning and recognizing dynamic hand gestures based on the following three assumptions that are consistent with neural evidence: 1) In the mammalian visual system, visual scenes are analyzed in parallel by two separated channels [7]. The ventral channel processes form

[*] This work was partially supported by the DAAD German Academic Exchange Service (Kz:A/13/94748) - CASY Project (Cognitive Assistive Systems), by the CAPES Brazilian Federal Agency for the Support and Evaluation of Graduate Education (p.n.5951-13-5), and by the DFG German Research Foundation (grant #1247) - International Research Training Group CINACS (Cross-modal Interaction in Natural and Artificial Cognitive Systems).

[1] Kinect for Windows. `http://www.microsoft.com/en-us/kinectforwindows`
ASUS Xtion Pro Live. `http://www.asus.com/Multimedia/Motion_Sensor/Xtion_PRO_LIVE/`

ESANN 2014 proceedings, European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning. Bruges (Belgium), 23-25 April 2014, i6doc.com publ., ISBN 978-287419095-7. Available from http://www.i6doc.com/fr/livre/?GCOI=28001100432440.

features and recognizes shape snapshots while the dorsal channel recognizes location and motion properties in terms of optic-flow patterns [9]; 2) Both channels contain hierarchies to extrapolate shape and optic-flow features with increasing complexity [8][9], from low- to high-level representations of the visual stimuli; 3) Input-driven self-organization of visual information is crucial for the cortex to organize by tuning itself according to the distribution of the inputs [10].

Our framework is motivated by these three principles and composed of two main modules. In Section 2 we introduce the first module to estimate statistical hand shape and motion information from visual scenes. The second module consists of a neurally inspired architecture for the clustering in two parallel processing streams of the extracted visual cues and their combination. In Section 3 we describe the learning and recognition phase based on a hierarchy of self-organizing maps (SOM) that represent increasingly complex properties of gestures. In Section 3 we provide experimental results and an evaluation for the recognition of a set of learned gestures from RGB-D videos. In Section 4 we present our concluding remarks and future work.

2 Statistical Hand Action and Pose Estimation

In order to extrapolate the dynamics of gestures, we use our algorithm for Statistical Hand Action and Pose Estimation (SHAPE) that combines two techniques: one for extracting inter-frame motion characteristics and the other for estimating the shape of the hand. The visual result of the segmentation is depicted in Fig. 1. The main role of SHAPE is to filter spatiotemporal properties from color and depth cues to extract gesture information. This process represents a convenient trade-off for reducing the amount of processed information without losing the most relevant featural properties. The output of each processing stream is a set of multi-dimensional vectors representing hand shape and motion features subsequently used for clustering. To extract motion information we first segment the foreground depth data cloud corresponding to the hand. This allows for efficient estimations of hand position in real-world coordinates. At every frame i, we estimate the hand centroid $C = (c^x, c^y, c^z)$ of the data cloud as the mean position of all the points. We determine the hand velocity S_i as the inter-frame difference in pixels between C_i and C_{i-1}. We then estimate the relative direction of motion in space as $D_i = \{S_i^x/d,\ S_i^y/d,\ S_i^z/d\}$ with $d = \sqrt{(S_i^x)^2 + (S_i^y)^2 + (S_i^z)^2}$. This representation expresses the intensity of direction-selective motion between consecutive frames with respect to the position of the sensor.

For estimating hand poses we first extract color information and then obtain the hand contour. We use an extension of the convexity approach technique [11] that can describe a hand pose using dynamically selected points in the hand contour. For each hand posture the points selection is minimized so that the feature vector contains only the minimal number of features required for the pose representation. The Convexity Approach is invariant to rotation and scale, and it can be used to recognize gestures executed by different users. However, in order to better extrapolate motion properties of dynamic gestures, it may be convenient

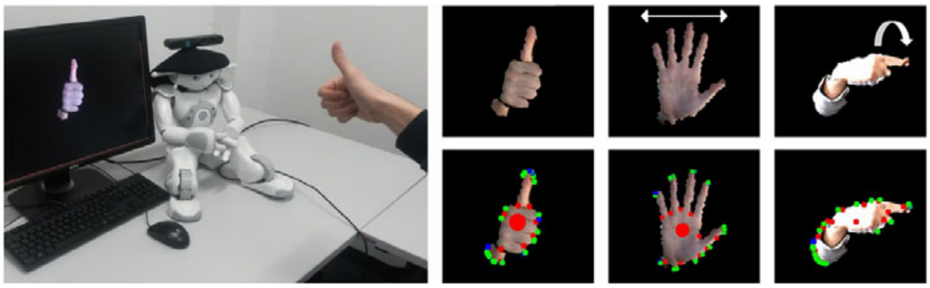

Fig. 1: Example of hand segmentation and pose estimation with SHAPE for static and dynamic gestures.

to introduce variance to rotation. We extract the contour centroid information using image moments [12], and then calculate the angle θ between a vertical line across the centroid and each point of the set. We use Hu moments [12] of the selected points, instead of their distances, for reducing the dimensionality of feature vectors. Hu moments are described in seven functions I using the central moment of image pixels. They are invariant to rotation and scale. Therefore, the hand pose is expressed in terms of a vector containing the set of Hu moments and the hand angle, formally expressed as $P_i = \{I_1, ..., I_7, \theta\}$.

3 Learning Framework

We propose a neurally inspired approach to process shape-motion features separately by two different streams. An overall overview of the learning framework is depicted in Fig. 2. First, spatiotemporal properties of the visual scene are processed by two parallel channels: one channel for processing shape features and the other channel for motion properties [7][9]. Second, inherent spatiotemporal dependencies of dynamic gestures can be implicitly learned by a hierarchical architecture to obtain increasingly abstract representations of the sensory inputs [8][9]. Therefore, both channels comprise hierarchies of competitive networks to extrapolate gesture dynamics from sequences of shape snapshots and optic-flow patterns respectively. Synchronized multi-cue features, i.e. features coming from the same time frames, are combined by a subsequent network for a more robust representation of the gesture. Our third assumption is based on cortical input-driven self-organization, suggesting that specific areas of the visual cortex are composed of topographically arranged neural structures that organize according to the distribution of the inputs [10]. Therefore, encoded representations can be learned with an unsupervised scheme by adaptively obtaining the feature subspace. The processing of multi-dimensional flow vectors with a self-organizing map (SOM) [14] has shown to be a plausible and efficient model for recognizing behavioral patterns [4][5][6]. A SOM network consists of an input and a competitive layer. Each unit j has an associated d-dimensional prototype vector $p_j = [p_{j,1}, p_{j,2}, ..., p_{j,d}]$. The SOM algorithm computes the models so that

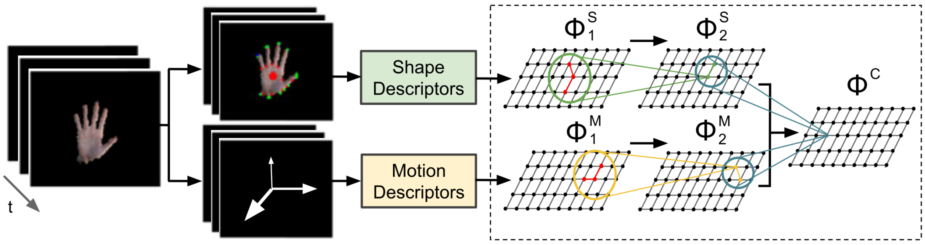

Fig. 2: FINGeR pipeline for the hierarchical SOM-based clustering of encoded gestures with the SHAPE algorithm. Pose and motion properties are processed by two different streams, Φ^S and Φ^M respectively. Synchronized multi-cue representations are subsequently combined by Φ^C.

they describe the domain of observations preserving the topological properties of the training data. For each input vector $x = (x_1, ..., x_d)$, the best matching unit (BMU) b for x_i is selected by the smallest Euclidean distance:

$$b(x_i) = argmin_j \|x_i - p_j\| \quad . \tag{1}$$

The hierarchical clustering is structured into two separate streams for the parallel processing of shape and motion cues, each composed of two SOM layers. We label training samples so that during the training labels are propagated along the hierarchy and are associated to prototype vectors. Each competitive network is trained with a batch variant of the SOM algorithm. Batch learning requires fewer parameters and converges much faster than the traditional stepwise recursive algorithm. The parameters for the SOM map structure and training algorithm are the same as in our previous work [6].

The inputs of the first layers Φ^S_1 and Φ^M_1 are the training sets of sequentially encoded features for shape and motion, denoted as S and M respectively. After this training phase, chains of labeled best matching units for ordered training sequences produce time varying trajectories on each network map. For a given network Φ and a training set X, we define the set of map trajectories as $T(\Phi, X) = \{b(x_{i-2}), b(x_{i-1}), b(x_i, \lambda(x_i)) : i \in [3, n(X)]\}$, where b is defined by Eq. 1, $\lambda(x_i)$ is the label associated to x_i and $n(X)$ is the number of elements in X. The networks Φ^S_2 and Φ^M_2 in the second layer are trained with $T^S = T(S, \Phi^S_1)$ and $T^M = T(M, \Phi^M_1)$ respectively. This step produces a mapping with gesture segments from consecutive samples. The third layer represents the integration of shape and motion features as BMU pairs of synchronized single-cue trajectories. Therefore, the network Φ^C is trained with the set of pairs

$$P = \{\langle b(T(\Phi^S_2, T^S_g)), b(T(\Phi^M_2, T^M_g))\rangle \ : g \in [1, w]\} \quad , \tag{2}$$

where w is the number of computed trajectories. After this final training step, we compute the set of labelled prototype vectors for multi-cue gesture pairs, formally expressed as $P = \{\langle p_g, \hat{\lambda}(p_g)\rangle\}$. At recognition time, extracted cues are processed separately through the hierarchies. For every three novel observations, we compute one testing trajectory for each stream. For each computed testing pair sample p_{g+1}, the recognition output is the label $\hat{\lambda}(p_{g+1})$.

4 Experimental Results

We captured RGB-D videos with an ASUS Xtion sensor at a constant frame rate of 30 Hz, a pixel resolution of 640x480, and an operation range from 0.8 to 2 meters. To reduce sensor noise, we computed the median value for every 5 measurements. Each second of video segmentation processed with the SHAPE algorithm extracted therefore 6 poses and 6 motion vectors. The system was trained with 10 different gesture classes. Each gesture was performed 10 times by three different individuals for a total of 300 training gestures.

To evaluate accuracy at recognition time, each gesture class was performed 30 times from varying sensor distances within the operation range. We run experiments on the 300 testing gestures with single-cue information, i.e. motion and shape, and their combination. The recognition result is based on the statistical mode of the last 3 output labels obtained from the network. A quantitative improvement on using multi-cue combination over single cues can be seen in Fig. 3. For our test set, the average accuracy increase on using combined cues over motion and shape inputs individually is 17% and 13% respectively. Furthermore, multi-cue combination has shown better results also compared to choosing the best result between single-cue approaches, with an average improvement of 10%. Additional experiments suggested that the significance of overall improvement introduced by cue integration is dependent on the dynamics of the training set, i.e. gestures relying more on motion, shape or both properties to be correctly represented. The extraction of cues and recognition were performed with perceptually irrelevant latency providing real time characteristics.

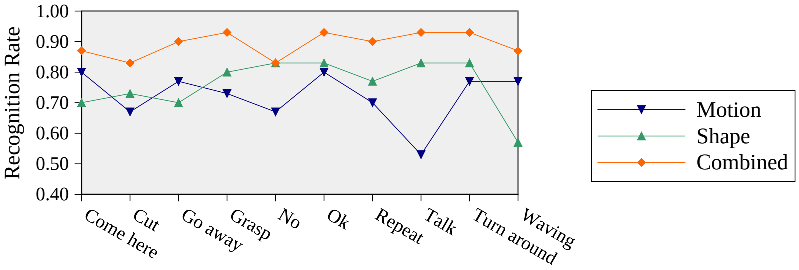

Fig. 3: Evaluation of FINGeR on a set of 10 dynamic gestures.

5 Conclusions and Future Work

We presented a modular two-stream framework for learning and recognizing multi-cue dynamic hand gestures based on three main assumptions that are consistent with neural evidence. The SHAPE algorithm extracts relevant statistical measurements of hand shape and motion from RGB-D videos to be encoded as small-dimensional flow vectors. The proposed hierarchical SOM-based clustering has shown to be a prominent method for learning dynamic gestures, providing perceptually real time recognition. Experimental results led to a quantitative enhancement of recognition rates for multi-cue combination over the single-cue strategy. The results obtained with FINGeR motivate further experiments with a wider number of more complex gestures, e.g. using both hands, and suggest that extensions of this approach may find a number of applications in the field of human-robot interaction: from a more natural communication with robots to the recognition of the sign language.

References

[1] J. Suarez and R. Murphy. *Hand gesture recognition with depth images: A review.* In *IEEE Intl. Symposium on Robot and Human Interactive Communication (RO-MAN)*, pages 411-417, France, 2012.

[2] Z. Ren, J. Meng and J. Yuan. *Depth camera based hand gesture recognition and its applications in human-computer interaction.* In *Intl. Conf. on Information, Communications, and Signal Processing*, pages 1-5, Singapore, 2012.

[3] R. Sun. *The Cambridge Handbook of Computational Psychology*, Cambridge University Press, New York, 2008.

[4] A. Shimada and R. Taniguchi. *Gesture recognition using sparse code of Hierarchical SOM.* In *Intl. Conf. on Pattern Recognition*, pages 1-4, Florida, US, 2008.

[5] W. Huang and Q.M.J Wu. *Human action recognition based on self organizing map.* In *Intl. Conf. on Acoustics Speech and Signal Processing*, pages 2130-2133, US, 2010.

[6] G. I. Parisi and S. Wermter. *Hierarchical SOM-based detection of novel behavior for 3D human tracking.* In *proceedings of the IEEE Intl. Joint Conf. on Neural Networks* (IJCNN), pages 1380-1387, US, 2013.

[7] G. Johansson. *Spatio-temporal differentiation and integration in visual motion perception. Psychology Research*, 38:379-393, 1976.

[8] M. Riesenhuber and T. Poggio. *Hierarchical models of object recognition in cortex*, *Nature Neuroscience*, 2(11):1019-1025, 1999.

[9] M. A. Giese and T. Poggio. *Neural mechanisms for the recognition of biological movements*, *Nature Reviews Neuroscience*, 4:179-192, 2003.

[10] R. Miikkulainen, J. A. Bednar, Y. Choe and J. Sirosh. *Computational Maps in the Visual Cortex*, Springer New York, 2005.

[11] P. V. A. Barros, N. T. M. Junior, J. M. M. Bisneto, B. J. T. Fernandes, B. L. D. Bezerra and S. M. M. Fernandes. *An effective dynamic gesture recognition system based on the feature vector reduction for SURF and LCS.* In *proceedings of the 23rd Intl. Conf. on Artificial Neural Networks (ICANN)*, pages 34-39, 2011.

[12] M-K. Hu. *Visual pattern recognition by moment invariants.* In *IRE Transactions on Information Theory*, 8(2):179-187, 1962.

[13] Y. Liu, Y. Yin, and S. Zhang. *Hand gesture recognition based on HU moments in interaction of virtual reality.* In *Intl. Conf. on Intelligent Human-Machine Systems and Cybernetics*, 1:145-148, 2012.

[14] T. Kohonen. *Self-organizing maps.* In *Series in Information Sciences*, Vol. 30, Springer Heidelberg, 1995.