HAND POSE RECOGNITION IN FIRST PERSON VISION THROUGH GRAPH SPECTRAL ANALYSIS

Mohamad Baydoun[1]*,Alejandro Betancourt*[1,2]*,Pietro Morerio*[1]*, Lucio Marcenaro*[1]*, Matthias Rauterberg*[2]*, Carlo Regazzoni*[1]

[1]Department of Electrical, Electronic, Telecommunications Engineering and Naval Architecture (DITEN) University of Genoa - Genoa, Italy.

[2]Department of Industrial Design. Eindhoven University of Technology. Eindhoven, Netherlands.

ABSTRACT

With the growing availability of wearable technology, video recording devices have become so intimately tied to individuals, that they are able to record the movements of users' hands, making hand-based applications one the most explored area in First Person Vision (FPV). In particular, hand pose recognition plays a fundamental role in tasks such as gesture and activity recognition, which in turn represent the base for developing human-machine interfaces or augmented reality applications. In this work we propose a graph-based representation of hands seen from the point of view of the user, obtained through the shape-fitting capability of a modified Instantaneous Topological Map. Spectral analysis of the graph Laplacian allows to arrange eigenvalues in vectors of features, which prove to be discriminative in classifying the considered hand poses.

Index Terms— First Person Vision, Egocentric Vision, Hand Pose, Graphs, Spectral Analysis

1. INTRODUCTION

The emergence of new wearable video recording devices such as action cameras and smart-glasses during the last five years has driven an important trend in the evolution of computer vision methods [1]. Namely, both the growing availability technology and the consequent huge production of Egocentric videos has increased the interest of researchers and computer scientist in developing methods to automatically process the data recorded from this new perspective.

Hand-related methods have quickly gained importance in FPV [2] since the spreading of wearable technology. Hands represent in fact one of the main interaction means with the surrounding environment [3] and thus also one of the most natural way to communicate with the device. Besides, most of people's activity do involve hands, which often perform gestures in the field of view of the users and of the body worn cameras [4]. Recognizing user's activity is believed to be essential in providing an augmented reality experience through the display of smart-glasses.

(a) Handwriting (b) Typing (c) Holding a cup

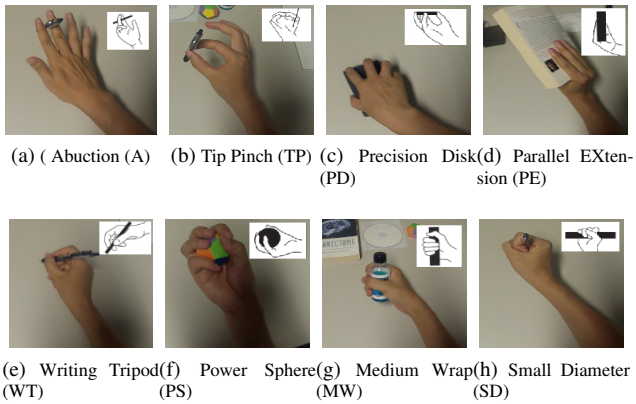
Fig. 1: Typical poses corresponding to the three different activities considered in our preliminary results (E1).

The problem of automatically recognizing hand poses has only recently been investigated in FPV by Kitani et al. [5]. The same authors propose a method to understand the functionality of human hands by analysing grasps poses using state of the art computer vision techniques [6]. In addition, their research exploits object recognition and hands-objects interaction analysis to deeper infer on users' activities [7]. Although the problem has already been addressed from third-person viewpoint, to the best of our knowledge the latter is the first attempt to analyse poses from the first-person perspective, even though only limited to grasp poses.

In this work, we propose a hand pose recognition method which exploits the property of graph-signals to encode shape information of objects. Representing objects with a graph is an idea which has become established with the growing interest in graph signal processing in the last few years. Such representation is for instance used in visual tracking, where tracked features are encoded in a graph, whose tracking is accomplished by exploiting graph matching techniques [8] from one frame to the following. It is indeed thanks to the growing interest in graph signal processing that techniques such as graph spectral analysis have been recently re-discovered. Just to mention, [9] uses graph spectral analysis for identifying properties of dynamically allocated data structures, while [10] even propose an extension the Nyquist-Shannon theory of sampling to signals defined on arbitrary graphs.

It is a common understanding that gesture recognition methods should be based on a first segmentation step in order to separate the hand region from the background [11]. In the field of FPV both pixel-by-pixel [12, 13, 14] and superpixel-

This work was partially supported by the Erasmus Mundus joint Doctorate in Interactive and Cognitive Environments, which is funded by the EACEA, Agency of the European Commission under EMJD ICE.



(a) (Abuction (A) (b) Tip Pinch (TP) (c) Precision Disk (PD) (d) Parallel EXtension (PE)

(e) Writing Tripod (WT) (f) Power Sphere (PS) (g) Medium Wrap (MW) (h) Small Diameter (SD)

Fig. 2: Eight grasps of the UTC dataset [6] (masks are provided), corresponding to Eight different taxonomic categories [17] used in E2.

based methods [15, 16] have been exploited to this end. In this work we employ a pixel-by-pixel approach, since single pixels will be used as the input of a Instantaneous Topological Map (ITM) to construct a graph representation of hands, as it will be detailed in section 2.2.

2. PROPOSED METHOD

The proposed method follows the work-flow depicted in Figure 3. After an image is acquired, a colour segmentation step outputs a binary image of the segmented hand. Such image is given as input to an artificial neural network, which outputs a graph whose topology mirrors the shape of the hands. The graph Laplacian is extracted and diagonalized, in order to compute its eigenvalues. A vector of ordered eigenvalues is then produced and given as input to SVM classifiers. In the following we will go into the details of each processing step.

2.1. Hand-segmentation

Input: Colour frame
Output: Binary image with hand segmentation

The aim here is to separate the foreground image represented by the hand region from the background. Hand-segmentation methods are commonly faced using the motion cues [18] or the colour and texture under a classification strategy [11, 12, 13]. Recent advances highlight colour-based strategies as the more reliable. Besides, it has been argued in [12] that being wearable devices personal (precisely as smartphones are), they can be trained specifically on the skin shade of a particular user. Our strategy relies on colour-based segmentation.

In particular we exploit the segmentation rule based on thresholds as proposed in [19], but we remark that the choice of the segmentation method is not the key aspect. For example, many approaches put great stress on smooth contours, which are not needed here. The aim of this level is simply to extract as many pixels as possible from the hand region, but even a poor segmenter allows in the following step to have a good hand representation, namely to construct a graph which is representative of the general hand pose. Any method with low false positives rates could be used expecting similar results.

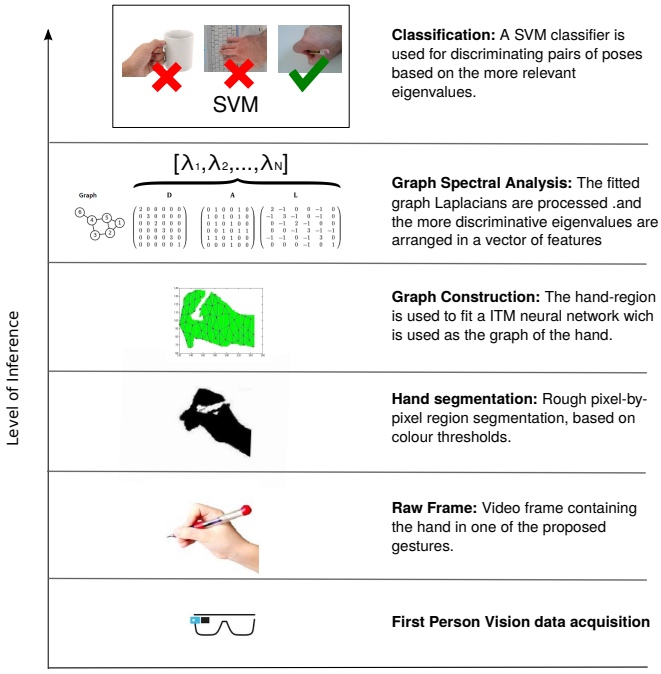


Fig. 3: Work-flow diagram of the proposed method.

The skin segmentation rules are devised in the $YCbCr$ colour space as suggested by both [19] and [12] and can be summarized as follows:

$$\begin{cases} Cb(i,j) \in R_{Cb}) \cap (Cr(i,j) \in R_{Cr}) \Rightarrow (i,j) \in hand, \\ D_1 \cup D_2 \cup D_3 \cup D_4 < T \Rightarrow (i,j) \in hand \end{cases}$$

where T is a threshold and D_1, D_2 are the euclidean distances between $Cb(i,j)$ and the upper and lower bounds of the range R_{Cb} and D_3, D_4 are the equivalent for the Cr channel.

2.2. Graph construction

Input: Binary mask image
Output: Graph representing the hand shape

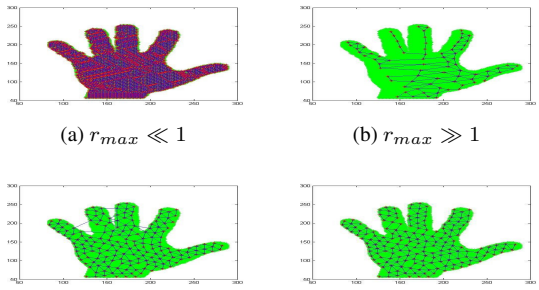
A graph is a structure composed by nodes connected by edges. In this work we consider non-directed graphs and take for granted that the reader has basic knowledge about them.

As already mentioned, graphs has recently been employed for representing visual objects to be tracked [8]. The main idea is there that objects can be represented as an ensemble

of sub-parts (nodes), one related to the other (edges). In this level we propose to encode the hand-shape as a graph structure, which is assumed to capture enough discriminative information to differentiate between poses. This block takes care of the constructions of such graph, exploiting the fitting capabilities of a neural network, namely a modified version of the Instantaneous Topological Map (ITM) [20].

The network is fed with the pixels from the hand mask, which fire neurons making them adapt to the stimuli. Links are modified accordingly. The main difference from the standard ITM lies in the fact that the map is able to fit concave shapes, since links which do cross white regions are removed Figure 4(d). Such links may appear when the parameter r_{max} is larger than a concavity, as it can be noticed from Figure 4(b). However, a very low r_{max} results in a very large number of nodes as in the case of 4(a). This is not desirable since it will produce a huge Laplacian matrix to be processed, as discussed in the next subsection. In our case the value is fixed to $r_{max} = 10$ in order to have between 50 and 150 nodes Figure 4(c). It is important to note that the number of nodes can change between different images but they will be of the same order of magnitude. For what concern ε, this parameter is tuned heuristically and its value is fixed to $\varepsilon = 0.1$. For a detailed discussion on the tuning of the two parameters the reader can refer to [21].

In terms of computational complexity, the matching step scales with the number of neurons, which is implicitly controlled by the parameter r_{max}.



(a) $r_{max} \ll 1$ (b) $r_{max} \gg 1$

(c) optimal value of $r_{max} \approx 10$ (d) external links removed

Fig. 4: Some examples of the adjusted graph trough the modified ITM algorithm (typing pose)

2.3. Graph spectral analysis

Input: Graph
Output: Vector of features

It is very common to deal with a graph in one of its matrix representation. The one we use in this work is the normalized version of the Laplacian, given by $L = I - D^{-\frac{1}{2}} A D^{-\frac{1}{2}}$, where D is the degree matrix (number of connection of each node on the diagonal) and A is the adjacency matrix (1 where a link exist, 0 elsewhere). Graph theory shows that the eigenvalues λ_i of L (which are real, being it symmetric) if sorted in ascending order can provide valuable information about the structure of the graph and will have interesting properties: i) they lie in the range $[0 : 2]$, ii) $\lambda_0 = 0$ with a multiplicity equal to the number of connected components of the graph, iii) λ_1 carries information about the general connectivity of graph. The other eigenvalues do not have semantic meaning, but they encode valuable information about the graph structure. This is why, starting from λ_1, we consider a vector of Laplacian eigenvalues as features to discriminate among hands poses.

2.4. Classification

Input: Vector of features
Output: Pose id

The top level of our structure is the classification step. For this purpose we use the vector of eigenvalues to train a Support Vector Machine (SVM). In the results section we initially test our approach as a pair-wise classifier to validate the discriminative power of the graph representation. This experiment uses our own dataset and the results are denoted as (E1). As an extension to these results we use 8 gestures of the UTC dataset under a multi-class approach. A more applicable method must be trained using more gestures.

3. RESULTS AND DISCUSSION

3.1. Preliminary investigation (E1)

Three different gestures with three typical poses are considered for a preliminary investigation. Samples of the employed images are shown in Figure 1. For simplicity we will refer to them as to pose 1, 2 and 3 respectively. A set of 40 heterogeneous pictures was collected for each pose. After extracting graph eigenvalues from each image, as illustrated in the previous section, poses were compared pair by pair, in a *k*-fold validation framework (more precisely we adopted the leave-one-out scheme). Classification true positive rates for different number of eigenvalues are reported in Figure 5(a). Note that using only the first non-zero eigenvalue yields to poor results, since the number of connections within the graphs are similar for all the gestures. However, by adding more eigenvalues the performances increase and stabilize around 0.9 in all the cases.

It is noteworthy that the case of *Gesture 1 vs 3* needs more eigenvalues to reach the maximum performance. This is easily explained by the fact that the two poses are similar, being the hand closed in both the cases. Tables 1(a)(b)(c) show the confusion matrices for the three experiments. As expected, the most unbalanced and challenging case is *Pose 1 vs Pose 3*.

For what concerns the complexity of the method, the most time consuming step is graph construction, as discussed in subsection 2.2: it strongly depends on the number of nodes, to be controlled by r_{max}. Such a number also influences the complexity of the subsequent matrix diagonalization.

Table 1: Pairwise confusion of the classification accuracy with 10 eigenvalues (E1).

(a) Pose 1 vs Pose 2

		SVM	
		Pose 1	Pose 2
Value	Pose 1	0.85	0.15
	Pose 2	0.05	0.95

(b) Pose 1 vs Pose 3

		SVM	
		Pose 1	Pose 3
Value	Pose 1	0.80	0.20
	Pose 3	0.00	1.00

(c) Pose 2 vs Pose 3

		SVM	
		Pose 2	Pose 3
Value	Pose 2	0.95	0.05
	Pose 3	0.15	0.85

Table 2: Confusion matrices of the classification accuracy (E2).

(a) 1 eigenvalues

		SVM							
		A	TP	PD	PE	WT	PS	MW	SD
Value	A	0.20	0.20	0.20	0.40	0.00	0.00	0.00	0.00
	TP	0.25	0.45	0.10	0.15	0.00	0.00	0.00	0.00
	PD	0.05	0.20	0.30	0.05	0.15	0.05	0.20	0.00
	PE	0.20	0.25	0.15	0.30	0.05	0.025	0.00	0.025
	WT	0.00	0.00	0.40	0.15	0.10	0.05	0.25	0.05
	PS	0.10	0.025	0.30	0.10	0.10	0.025	0.20	0.15
	MW	0.05	0.00	0.20	0.00	0.10	0.05	0.30	0.30
	SD	0.00	0.00	0.15	0.00	0.10	0.15	0.425	0.175

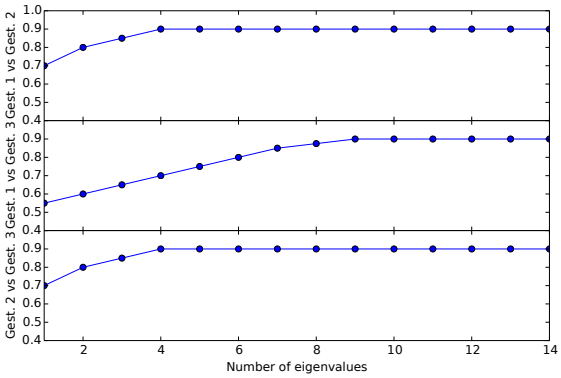
(b) 9 eigenvalues

		SVM							
		A	TP	PD	PE	WT	PS	MW	SD
Value	A	0.85	0.05	0.05	0.05	0.00	0.00	0.00	0.00
	TP	0.075	0.85	0.025	0.025	0.00	0.00	0.00	0.00
	PD	0.15	0.05	0.75	0.05	0.00	0.00	0.00	0.00
	PE	0.05	0.05	0.025	0.85	0.025	0.00	0.00	0.00
	WT	0.00	0.00	0.05	0.05	0.80	0.05	0.05	0.00
	PS	0.00	0.00	0.00	0.05	0.10	0.70	0.10	0.05
	MW	0.00	0.00	0.00	0.00	0.025	0.025	0.80	0.15
	SD	0.00	0.00	0.00	0.00	0.025	0.025	0.10	0.85

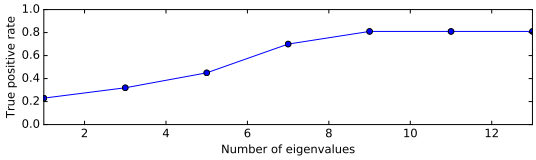
3.2. UTC dataset (E2)

Given the encouraging results of the preliminary investigation, we proceeded to validate our findings on a more structured public dataset [6]. For each of the 3 categories proposed in [17], we chose some kind of grasp (Fig. 2). Since the UTC dataset provides hand masks, the segmentation step can be avoided. For the proposed gestures a multi-class SVM [22] is trained while increasing the number of eigenvalues where the results are presented in Fig. 5(b). Note that for this particular dataset the accuracy stabilizes when 9 eigenvalues are used.

Finally, the confusing matrices for the two different eigenvalues of the 8 UTC gestures are presented in Table 2. The results are obtained by following a leave one out cross validation. As expected, while the use of 1 eigenvalue makes the decision impossible to differentiate the gestures, the 9 eigenvalues achieve a good classification score. However, the confusion remains high in case the gestures are similar (for example MW, SD). Besides, as the PS share approximately the same eigenvalues with WT, MW and SD, the PS proves to be the most difficult pose to differentiate compared to others.



(a) E1



(b) E2

Fig. 5: The effect of the number of eigenvalues on the accuracy of the classifiers in the two experiments.

4. CONCLUSION

In this work we have shown how a graph representation of hand shapes can be exploited in a pose recognition problem. Vectors of graph Laplacian eigenvalues proved to be robust features in discriminating pairs of poses. In the case of similar poses, more information is need for reaching an optimal classification accuracy. Results are encouraging although far from real time, which can be approached by optimizing implementation.

Still, a more applicable method should be tested on a structured multi-class approach using more gestures. Another future research may include an analysis of the robustness of the eigenvalue representation against segmentation noise.

5. REFERENCES

[1] A Betancourt, P Morerio, C.S. Regazzoni, and M Rauterberg, "The Evolution of First Person Vision Methods: A Survey," *IEEE Transactions on Circuits and Systems for Video Technology*, pp. 1–1, 2015.

[2] A Betancourt, P Morerio, L Marcenaro, E Barakova, M Rauterberg, and C.S. Regazzoni, "Towards a Unified Framework for Hand-based Methods in First Person Vision," in *IEEE International Conference on Multimedia and Expo*, Turin, 2015, IEEE.

[3] A Fathi, A Farhadi, and J Rehg, "Understanding Egocentric Activities," in *International Conference on Computer Vision*. Nov. 2011, pp. 407–414, IEEE.

[4] Alejandro Betancourt, Pietro Morerio, Emilia I Barakova, Lucio Marcenaro, Matthias Rauterberg, and Carlo S Regazzoni, "A dynamic approach and a new dataset for hand-detection in first person vision.," in *International conference on Computer Analysis of Images and Patterns*, 2015.

[5] Minjie Cai, Kris M Kitani, and Yoichi Sato, "A scalable approach for understanding the visual structures of hand grasps," in *Robotics and Automation (ICRA), 2015 IEEE International Conference on*. IEEE, 2015, pp. 1360–1366.

[6] De-An Huang, Wei-Chiu Ma, Minghuang Ma, and Kris M. Kitani, "How do we use our hands? discovering a diverse set of common grasps.," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2015.

[7] Tatsuya Ishihara, Kris Kitani, Wei-Chiu Ma, Hironobu Takagi, and Chieko Asakawa, "Recognizing hand-object interactions in wearable camera videos," in *International Conference on Image Processing*, 2015.

[8] Zhaowei Cai, Longyin Wen, Zhen Lei, N. Vasconcelos, and S.Z. Li, "Robust deformable and occluded object tracking with dynamic graph," *Image Processing, IEEE Transactions on*, vol. 23, no. 12, pp. 5497–5509, Dec 2014.

[9] M.Z. Malik and S. Khurshid, "Dynamic shape analysis using spectral graph properties," in *Software Testing, Verification and Validation (ICST), 2012 IEEE Fifth International Conference on*, April 2012, pp. 211–220.

[10] A. Anis, A. Gadde, and A. Ortega, "Towards a sampling theorem for signals on arbitrary graphs," in *Acoustics, Speech and Signal Processing (ICASSP), 2014 IEEE International Conference on*, May 2014, pp. 3864–3868.

[11] C Li and K Kitani, "Model Recommendation with Virtual Probes for Egocentric Hand Detection," in *ICCV 2013*, Sydney, 2013, IEEE Computer Society.

[12] P Morerio, L Marcenaro, and C Regazzoni, "Hand Detection in First Person Vision," in *Information Fusion*, Istanbul, 2013, University of Genoa, pp. 1502 – 1507.

[13] C Li and K Kitani, "Pixel-Level Hand Detection in Egocentric Videos," in *Computer Vision and Pattern Recognition*. June 2013, pp. 3570–3577, Ieee.

[14] Alejandro Betancourt, Pietro Morerio, Emilia I. Barakova, Lucio Marcenaro, Matthias Rauterberg, and Carlo S. Regazzoni, "Left/right hand segmentation in egocentric videos," *CoRR*, vol. abs/1607.06264, 2016.

[15] G Serra, M Camurri, and L Baraldi, "Hand Segmentation for Gesture Recognition in Ego-vision," in *Workshop on Interactive Multimedia on Mobile & Portable Devices*, New York, NY, USA, 2013, pp. 31–36, ACM Press.

[16] Pietro Morerio, Gabriel Claudiu Georgiu, Lucio Marcenaro, and Carlo Regazzoni, "Optimizing Superpixel Clustering for Real-Time Egocentric-Vision Applications," *IEEE Signal Processing Letters*, vol. 22, no. 4, pp. 469–473, Apr. 2015.

[17] I.M. Bullock, T. Feix, and A.M. Dollar, "Finding small, versatile sets of human grasps to span common objects," 2013, pp. 1068–1075.

[18] S Sundaram and W Cuevas, "High Level Activity Recognition Using Low Resolution Wearable Vision," *Computer Vision and Pattern Recognition Workshops*, pp. 25–32, June 2009.

[19] E. Stergiopoulou and N. Papamarkos, "Hand gesture recognition using a neural network shape fitting technique," *Engineering Applications of Artificial Intelligence*, vol. 22, no. 8, pp. 1141–1158, 2009, cited By 81.

[20] Jan Jockusch and Helge Ritter, "Instantaneous topological mapping model for correlated stimuli," 1999, vol. 1, pp. 529–534, cited By 34.

[21] P. Morerio, L. Marcenaro, and C.S. Regazzoni, "A generative superpixel method," 2014, cited By 1.

[22] Chih-Wei Hsu and Chih-Jen Lin, "A comparison of methods for multiclass support vector machines," *Neural Networks, IEEE Transactions on*, vol. 13, no. 2, pp. 415–425, Mar 2002.