Università degli Studi di Modena e Reggio Emilia

Dipartimento di Ingegneria "Enzo Ferrari"

Corso di Laurea Magistrale in Ingegneria Informatica

Hand Gesture recognition in Ego-Centric videos

Candidato:
Lorenzo Baraldi

Relatore:
Prof. Rita Cucchiara
Correlatore:
Ing. Giuseppe Serra

Anno Accademico 2013-2014

To Emanuela

Acknowledgements

Foremost, I would like to express my sincere gratitude to my advisors Prof. Rita Cucchiara and Dr. Giuseppe Serra for the continuous support of my study, for their patience, motivation, enthusiasm, and knowledge. Their guidance helped me in all the time of research and writing of this thesis. I could not have imagined having better advisors for my study.

Besides my advisors, I would like to thank the rest of the Imagelab research group: Costantino Grana, Simone Calderara, Paolo Santinelli, Michele Fornaciari, Marco Manfredi, Franesco Solera, Martino Lombardi and Patrizia Varini, for their encouragement, insightful comments, and hard questions. I am also grateful to Prof. Greg Mori, for enlightening me the first glance of research during a visit in Modena.

Special thanks go to my labmate Francesco Paci, for the stimulating discussions, for the days we were working together before deadlines, and for all the fun we have had in the last months. Also I thank my friends Chiara Ferrari, Emanuele Benatti, Davide Setti, and in particular Michela Benedetti, who encouraged me to study ego-centric vision. Without her, this thesis would not probably exists.

Last but not the least, I would like to thank my parents Franco and Elisabetta, for giving birth to me at the first place and supporting me throughout my life, and my sister Alessia.

Contents

List of Figures

List of Tables

Chapter 1

Wearable devices and new human-machine interfaces: an overview

1.1 Where to wear an extra eye?

Positioning an optical device on the human body is quite a problematic task, as occlusion, motion, social issues as well as criteria related to the purpose of the device must be taken into account.[1]. Following the work of Mayol *et al.* [1], in this section we give a detailed overview on the best places where to put a wearable camera.

Cameras used for wearable applications fall into two categories for this discussion; static narrow-view devices and omnidirectional devices. Omnidirectional devices include catadioptric, fish-eye and active systems where either the entire field-of-viewis imaged at low resolution, or in the active case the high-resolution narrow-view sensor moved to any orientation. Narrow-view static cameras can only ever see a small part of the user or their environment, and placement is therefore entirely driven by the task. For wide-angle or omnidirectional sensors placement is less constrained and a range of positions are possible.

1.2 3D cameras

1.3 In Closing

[1] http://www.robots.ox.ac.uk/~wmayol/3D/nancy_matlab.html

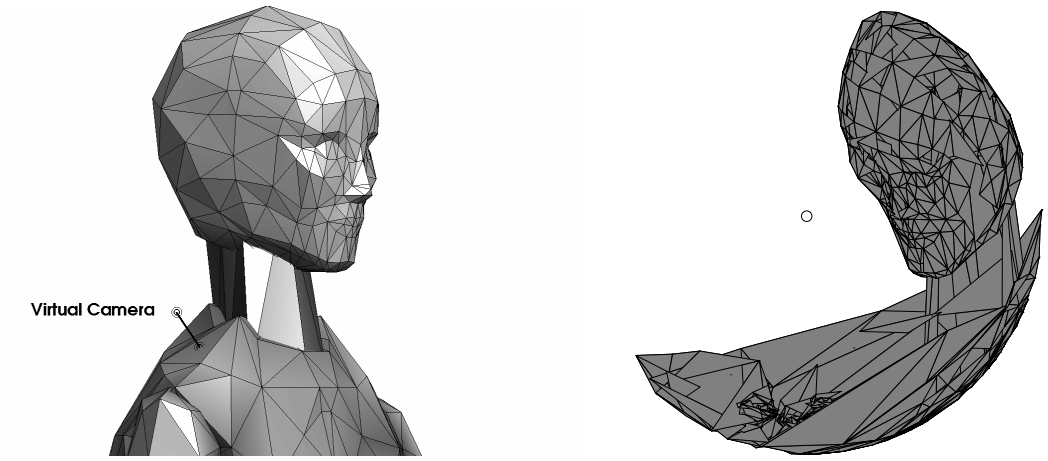

FIGURE 1.1: An electron (artist's impression).

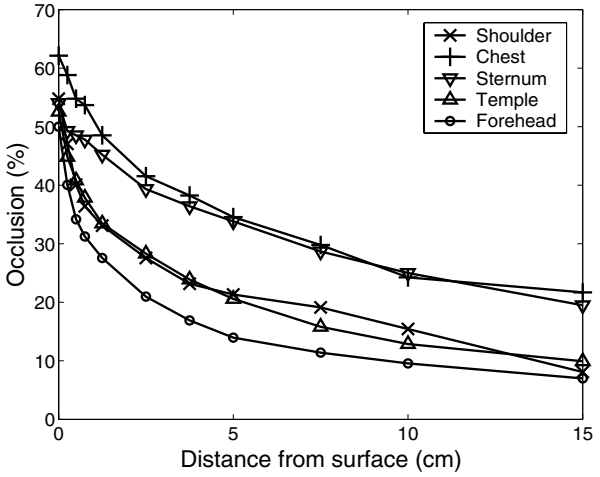

FIGURE 1.2: An electron (artist's impression).

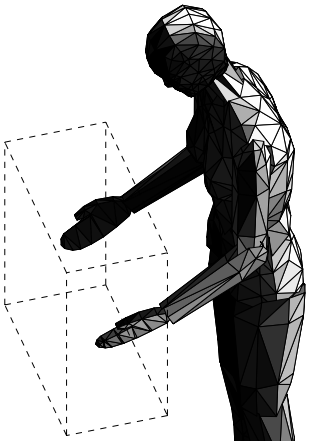

FIGURE 1.3: An electron (artist's impression).

Chapter 2

Hand segmentation in ego-centric videos

As stated before, a hand segmentation mask is used to distinguish between camera and hand motions, and to prune away all the trajectories that do not belong to the user's hand. In this way, our descriptor captures hands movement and shape as if the camera was fixed, and disregards the noise coming from other moving regions that could be in the scene.

At each frame we extract superpixels using the SLIC algorithm [**?**], that performs a k-means-based local clustering of pixels in a 5-dimensional space, where color and pixel coordinates are used. Superpixels are then represented with several features: histograms in the HSV and LAB color spaces (that have been proven to be good features for skin representation [**?**]), Gabor filters and a simple histogram of gradients, to discriminate between objects with a similar color distribution.

2.0.0.1 Illumination invariance

To deal with different illumination conditions we train a collection of Random Forest classifiers indexed by a global HSV histogram, instead of using a single classifier. Hence, training images are distributed among the classifiers by a k-means clustering on the feature space. By using a histogram over all three channels of the HSV color space, each scene cluster encodes both the appearance of the scene and its illumination. Intuitively, this models the fact that hands viewed under similar global appearance will share a similar distribution in the feature space. Given a feature vector $\mathbf{l}$ of a superpixel $\mathbf{s}$ and a global appearance feature $\mathbf{g}$, the posterior distribution of $\mathbf{s}$ is computed by marginalizing over different scenes c:

$$P(\mathbf{s}|\mathbf{l},\mathbf{g}) = \sum_{c} P(\mathbf{s}|\mathbf{l},c)P(c|\mathbf{g}) \tag{2.1}$$

where $P(\mathbf{s}|\mathbf{l},c)$ is the output of a global appearance-specific classifier and $P(c|\mathbf{g})$ is a conditional distribution of a scene c given a global appearance feature $\mathbf{g}$. In test phase, the conditional $P(c|\mathbf{g})$ is approximated using an uniform distribution over the five nearest models learned at training. It is important to highlight that the optimal number of classifiers depends on the characteristics of the dataset: a training dataset with several different illumination conditions, taken both inside and outside, will need an higher number of classifiers than one taken indoor. In addition, we model the hand appearance not only considering illumination variations, but also including semantic coherence in time and space.

2.0.0.2 Temporal coherence

To improve the foreground prediction of a pixel in a frame, we replace it with a weighted combination of its previous frames, since past frames should affect the prediction for the current frame.

We define a smoothing filter for a pixel x_t^i from frame t as:

$$\begin{aligned} P(x_t^i \;\; = \;\; 1) = \sum_{k=0}^{d} w_k (P(x_t^i = 1|x_{t-k}^i = 1) \cdot \\ \cdot P(x_{t-k}^i = 1|\mathbf{l}_{t-k},\mathbf{g}_{t-k}) + P(x_t^i = 1|x_{t-k}^i = 0) \\ \cdot P(x_{t-k}^i = 0|\mathbf{l}_{t-k},\mathbf{g}_{t-k})) \end{aligned} \tag{2.2}$$

where d is the number of past frames used, and $P(x_{t-k}^i = 1|\mathbf{l}_{t-k},\mathbf{g}_{t-k})$ is the probability that a pixel in frame $t-k$ is marked as hand part, equal to $P(\mathbf{s}|\mathbf{l}_{t-k},\mathbf{g}_{t-k})$, being x_t^i part of $\mathbf{s}$. In the same way, $P(x_{t-k}^i = 0|\mathbf{l}_{t-k},\mathbf{g}_{t-k})$ is defined as $1 - P(\mathbf{s}|\mathbf{l}_{t-k},\mathbf{g}_{t-k})$. Last, $P(x_t^i = 1|x_{t-k}^i = 1)$ and $P(x_t^i = 1|x_{t-k}^i = 0)$ are prior probabilities estimated from the training set as follows:

$$P(x_t^i = 1|x_{t-k}^i = 1) = \frac{\#(x_t^i = 1, x_{t-k}^i = 1)}{\#(x_{t-k}^i = 1)}$$

$$P(x_t^i = 1|x_{t-k}^i = 0) = \frac{\#(x_t^i = 1, x_{t-k}^i = 0)}{\#(x_{t-k}^i = 0)} \tag{2.3}$$

where $\#(x^i_{t-k} = 1)$ and $\#(x^i_{t-k} = 0)$ are the number of times in which x^i_{t-k} belongs or not to a hand region, respectively; $\#(x^i_t = 1, x^i_{t-k} = 1)$ is the number of times that two pixels at the same location in frame t and $t - k$ belong to a hand part; similary $\#(x^i_t = 1, x^i_{t-k} = 0)$ is the number of times that a pixel in frame t belongs to a hand part and the pixel in the same position in frame $t - k$ does not belong to a hand region.

2.0.0.3 Spatial consistency

Given pixels elaborated by the previous steps, we want to exploit spatial consistency to prune away small and isolated pixel groups that are unlikely to be part of hand regions and also aggregate bigger connected pixel groups. For every pixel x, we extract its posterior probability $P(x^t_i)$ and use it as input for the GrabCut algorithm [**?**]. Each pixel with $P(x^t_i) \geq 0.5$ is marked as foreground, otherwise it's considered as part of background. After the segmentation step, we discard all the small isolated regions that have an area of less than 5% of the frame and we keep only the three largest connected components.

2.1 Experimental results

Chapter 3

Towards ego-vision human-machine interfaces: gesture recognition

3.1 A distributed network of smart sensors to improve training

3.2 Support Vector Machines Hidden Markov Models

3.3 Experimental results

Chapter 4

Conclusion

4.1 Publications

Appendix A

Appendix Title Here

Write your Appendix content here.

Bibliography

[1] WW Mayol, B Tordoff, and DW Murray. On the positioning of wearable optical devices. Technical report, Technical Report OUEL224101. Oxford University, 2001.