PERSONALIZING A SMARTWATCH-BASED GESTURE INTERFACE WITH TRANSFER LEARNING

Gabriele Costante[*], *Lorenzo Porzi*[*†], *Oswald Lanz*[†], *Paolo Valigi*[*], *Elisa Ricci*[*†]

[*] University of Perugia, Perugia, Italy [†] Fondazione Bruno Kessler, Trento, Italy

ABSTRACT

The widespread adoption of mobile devices has lead to an increased interest toward smartphone-based solutions for supporting visually impaired users. Unfortunately the touch-based interaction paradigm commonly adopted on most devices is not convenient for these users, motivating the study of different interaction technologies. In this paper, following up on our previous work, we consider a system where a smartwatch is exploited to provide hands-free interaction through arm gestures with an assistive application running on a smartphone. In particular we focus on the task of effortlessly customizing the gesture recognition system with new gestures specified by the user. To address this problem we propose an approach based on a novel transfer metric learning algorithm, which exploits prior knowledge about a predefined set of gestures to improve the recognition of user-defined ones, while requiring only few novel training samples. The effectiveness of the proposed method is demonstrated through an extensive experimental evaluation.

Index Terms— Gesture recognition, smartwatch, transfer learning, Haar features, visual impairments.

1. INTRODUCTION

With their ever increasing computational power, low cost and widespread adoption, smartphones are arguably ideal targets for developing applications aimed at assisting disabled users in carrying out their daily activities. In particular, low-vision users could greatly benefit from applications that exploit the smartphone's camera to guide them and to help them avoiding potential dangers. However, the typical touch-based interfaces adopted on smart devices are not accessible for low-vision people, and traditional systems based on voice are generally not reliable enough when used in noisy environments.

Starting from these premises, in our previous work [1] we proposed a system where a smartphone hanging from the user's neck provides an assistive application controlled by arm gestures captured by a smartwatch. Expanding on this work, we tackle the problem of system reconfigurability, that is allowing the user to define his own set of gestures. In general, a desirable feature of this kind of system is that of only requiring the user to input a small number of sample gestures to reconfigure it. More so, in the case of systems designed for assistive tasks the customization process has to be as effortless as possible. Several previous works describe gesture recognition algorithms having this property [2, 3, 4]. However, to our knowledge none have considered the possibility of exploiting prior information to improve the recognition of user-defined gestures.

In this paper we propose a novel approach to reconfigurable gesture recognition based on Supervised Local Distance Learning. Prior knowledge about a set of predefined gestures is exploited to improve the recognition of user-defined gestures through a novel domain adaptation technique. Only a small number of gesture repetitions is required from the user to reconfigure the system: with two repetitions per gesture we obtain an average accuracy of 85% on a set of six target gestures and the accuracy rises to 93% if three repetitions are available. While the training phase requires some computations on a remote server (which are only performed once offline), the recognition can be performed in real time on a standard smartphone.

2. RELATED WORKS

Several previous works have addressed the topic of making portable devices more accessible for the visually impaired [5, 6]. Of particular relevance is Freevox [1], a smartwatch-like device designed with accessibility as a primary requirement. Differently from our system, the user interacts with Freevox using voice commands and a simplified touch interface. The use of smartwatches as gesture-based input devices has been probably first considered in the work of Bieber *et al.* [7], which underlines the fundamental distinction between the two tasks of gesture recognition and activity recognition. Common of several publications about smartwatch-based interaction is the adoption of custom or semi-custom devices. Morganti *et al.* [8] propose a wrist-worn device capable of detecting hand and finger gestures through flexible force sensors, while Bonino *et al.* [9] adapt a commercial device to be used for domotic applications developing a custom firmware. In contrast we choose to focus on a low cost device available off-the-shelf.

[1] http://myfreevox.com/en/

Gesture recognition from accelerometer data is generally treated as a classification problem, with different authors proposing different machine learning approaches to its solution. In particular, Support Vector Machines (SVM) [10, 11, 12], Hidden Markov Models (HMMs) [13, 2] and Bayesian Networks [14] have proven to be well suited to the task. Methods based on generative approaches like HMMs are generally limited in their applicability due to their computational complexity. SVMs on the other side usually offer lower computational requirements at classification time, making them preferable for real-time applications on low power devices. Of particular interest for our work are those algorithms targeted at recognizing user-defined gestures. Liu *et al.* [4] propose a system based on Dynamic Time Warping that can be trained with a single sample gesture and that keeps itself updated through template adaptation. Similarly the system proposed by Mantyjarvi *et al.* [2] can be trained with a single gesture and employs noise-distorted copies of that gesture to train a HMM. To our knowledge however no previous work exists on exploiting prior knowledge about a predefined set of gestures to improve the recognition of user-defined gestures.

Transfer learning [15] has recently gained great importance due to the increasing need of applications robust to changes of operating conditions. Transfer learning aims to improve a classification or regression model trained on few data (*i.e.* the target data), by exploiting knowledge from data of related tasks (source data). In particular, domain adaptation based on a distance learning framework has been previously proposed. In [16] a transfer metric learning method is introduced for visual object categorization. However, the same set of categories are assumed in the known and the novel domain. In [17] transfer learning with different categories in source and target data is addressed but, while we provide a local metric learning framework, they use a completely different approach based on spectral embedding. In general, we are not aware of a transfer learning method targeted to accelerometer-based gesture recognition applications.

3. SYSTEM OVERVIEW

In this work we consider a system composed of two devices: a smartphone which the visually impaired holds in front of his breast using a necklace and a wrist-worn smartwatch. The smartphone runs an application composed of several modules, each designed to assist the user carrying out a specific activity. Two of these modules have been described in [1]: a logo detector and a danger sign detector, while others (*e.g.* a vision-based assistant for navigating corridors) are currently being developed. The smartwatch acts as a remote acceleration sensor, providing a mean of capturing arm gestures. Each gesture is associated with an action to be performed on the system, *e.g.* switching between modules. We want to extend this system to allow the user to define his own set of personalized gestures. To this end we propose the transfer learning algorithm described in Section 4.

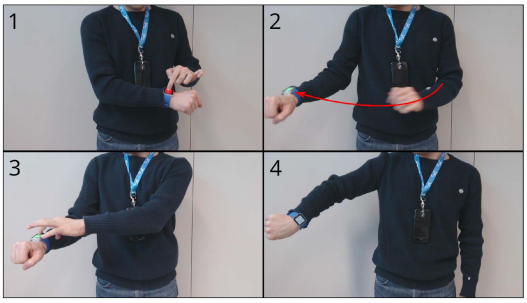


Fig. 1. A user interacting with the proposed system. 1) The gesture recognition is activated by tapping on the smartwatch. 2) The user performs a gesture. 3) The gesture's end is signaled with a second tap. 4) The system recognizes the gesture.

As mentioned in Section 1, the computation related to the gesture reconfiguration procedure is offloaded to a remote server due to its computational complexity, while the gesture recognition itself and the remaining application logic is entirely implemented on the smartphone. The two devices used in our testing are a Google Galaxy Nexus and a Sony SmartWatch™. This is a very low-cost smartwatch that can communicate wirelessly with any Android device trough a Bluetooth radio. Its integrated 3-axis accelerometer provides measurements at a sampling rate of 10 Hz.

A tap on the touch sensitive screen of the watch is used to signal the beginning and the end of a gesture. During this interval the system acquires a sequence of readings from the accelerometer, composed of one independent signal for each of the three axes of the sensor. As proposed in [10], we use the Haar Wavelet Transform to represent the acceleration signal. The first eight samples of the Haar transform of each channel are concatenated to form a 24-element vector that constitutes the input to the classification algorithm described in the next section. Figure 1 shows an user with the proposed system.

4. LEARNING ALGORITHM

In this section we present our transfer learning method for gesture recognition. The aim of the proposed approach is to train a distance metric to effectively classify a user-defined set of gestures, the *target set* $\mathcal{T}$, taking advantage of previous knowledge, *i.e.* a collection of gestures previously labeled (the *source set* $\mathcal{S}$). Since in our application we allow the user to record its own gestures, the source and the target set may contain, in general, different classes. We define $\mathcal{S} = \{(\mathbf{h}_1^s, y_1), (\mathbf{h}_2^s, y_2), \ldots, (\mathbf{h}_{N_s}^s, y_{N_s})\}$ and $\mathcal{T} = \{(\mathbf{h}_1^t, l_1), (\mathbf{h}_2^t, l_2), \ldots, (\mathbf{h}_{N_t}^t, l_{N_t})\}$, where $\mathbf{h}_i \in I\!R^d$ contains the Haar coefficients computed on the i-th gesture, $y_i, l_i \in I\!R$ are the source and the target labels and N_s and N_t are respectively the number of source and target data. As stated above, the source and the target sets may contain different classes, *i.e.* $y_i \in \mathcal{C}^s = \{C_1^s, \ldots C_{K_s}^s\}$,

Algorithm 1 Algorithm to solve (1)

Input: The sets $\mathcal{T}$ and $\mathcal{S}$, the regularization parameter λ, the number of iteration T, the threshold θ.

Compute $\mathcal{M}(C_i^s, C_j^t), \forall i, j$.
Initialize $\mathbf{W} = [\mathbf{w}_1, \mathbf{w}_2, \ldots, \mathbf{w}_{K_t}] = \mathbf{0}$.
for $t = 1, \ldots, T$ **do**
 Initialize $\mathbf{D} \in I\!R^{dK_t}$, $\mathbf{D} = \mathbf{0}$.
 $c = 0$.
 for $i, j, k = 1, \ldots, N_t$ **do**
  $\mathbf{X}_{ijk} = \mathbf{d}_{ik} - \mathbf{d}_{ij}$
  if $((1 - \mathbf{w}_{l_i}^T \mathbf{X}_{ijk} \geq 0) \wedge (l_k \neq l_i) \wedge (l_j = l_i))$
   $\mathbf{D}[(l_i - 1)d + 1 : l_i d] = \mathbf{D}[(l_i - 1)d + 1 : l_i d] + \mathbf{X}_{ijk}$
   $c = c + 1$.
  endif
 endfor
 for $q = 1, \ldots, N_s$ **do**
  Compute $p = \arg\min_{z \in \mathcal{C}^t} \mathcal{M}(y_q, z)$
  for $n, m = 1, \ldots, N_s$ **do**
   $\mathbf{X}_{qnm} = \mathbf{d}_{qn} - \mathbf{d}_{qm}$
   if $((1 - \mathbf{w}_p^T \mathbf{X}_{qnm} \geq 0) \wedge (y_n \neq y_q) \wedge (y_m = y_q))$
    Compute $r = \arg\min_{a \in \mathcal{C}^t, p \neq a} \mathcal{M}(y_n, a), y_n \neq y_q$
    if $(\mathcal{M}(y_q, p) \leq \theta \wedge \mathcal{M}(y_n, r) \leq \theta)$
     $\mathbf{D}[(p-1)d + 1 : pd] = \mathbf{D}[(p-1)d + 1 : pd] + \mathbf{X}_{qnm}$
     $c = c + 1$.
    endif
   endif
  endfor
 endfor
 $\boldsymbol{W}_{t+\frac{1}{3}} = \left(1 - \frac{1}{t}\right) \boldsymbol{W}_t + \frac{1}{c\lambda t}\mathbf{D}$
 $\boldsymbol{W}_{t+\frac{2}{3}} = \max\{0, \boldsymbol{W}_{t+\frac{1}{3}}\}$
 $\boldsymbol{W}_{t+1} = \min\{1, \frac{1}{\sqrt{\lambda}\|\boldsymbol{W}_{t+\frac{2}{3}}\|}\}\boldsymbol{W}_{t+\frac{2}{3}}$
endfor
Output: $\boldsymbol{W}$

$l_i \in \mathcal{C}^t = \{C_1^t, \ldots C_{K_t}^t\}$. Note that our framework can also handle the situation where $\mathcal{C}^t \cap \mathcal{C}^t = \emptyset$. In our scenario we allow a very short configuration phase where the user records 2-3 repetitions for each novel gesture class. It is clear that, having at disposal only very few training samples in the target set, it is very challenging to learn an effective recognition model. If one could take advantage of past knowledge, then a performance improvement could be achieved.

It is reasonable to suppose that not all the source data can be exploited in the target domain: we need to understand what information to transfer and what to discard. In our scenario this is achieved by evaluating the similarity between the source and the target classes. For this task we propose to use the Maximum Mean Discrepancy (MMD) [18] to compute the divergence between two classes C_i^s and C_j^t as follows:

$$\mathcal{M}^2(C_i^s, C_j^t) = \frac{1}{m_i^{s2}}\|K_{C_i^s}\|_1 - \frac{2}{m_i^s m_j^t}\|K_{C_i^s C_j^t}\|_1 + \frac{1}{m_j^{t2}}\|K_{C_j^t}\|_1$$

where K is the kernel matrix (in this paper we simply consider a linear kernel) and m_i^s and m_j^t are the number of samples of classes C_i^s and C_j^t.

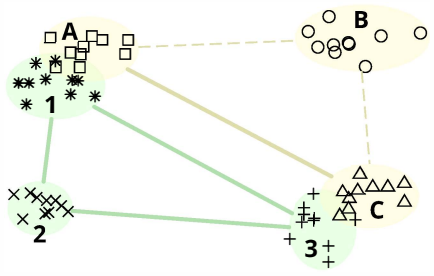


Fig. 2. A visual representation of our constraint selection strategy. A,B,C: source classes. 1,2,3: target classes. The constraints between A and C are preserved since A and C are similar to 1 and 3, those between A and B and between B and C are discarded as B is not similar to any of the target classes.

In our approach we first compute $\mathcal{M}(C_i^s, C_j^t)$ of every pair of source and target classes and then we learn a distance function to be used to classify novel gestures according to a Nearest Neighbor (NN) scheme. More specifically, we propose to learn a set of K_t distance functions $\delta_c(\mathbf{h}_i, \mathbf{h}_j) = \mathbf{w}_c^T \mathbf{d}(\mathbf{h}_i, \mathbf{h}_j) = \mathbf{w}_c^T \mathbf{d}_{ij} = \sum_n w_c^n (h_i^n - h_j^n)^2$, one for each target class. We learn them introducing the following optimization problem:

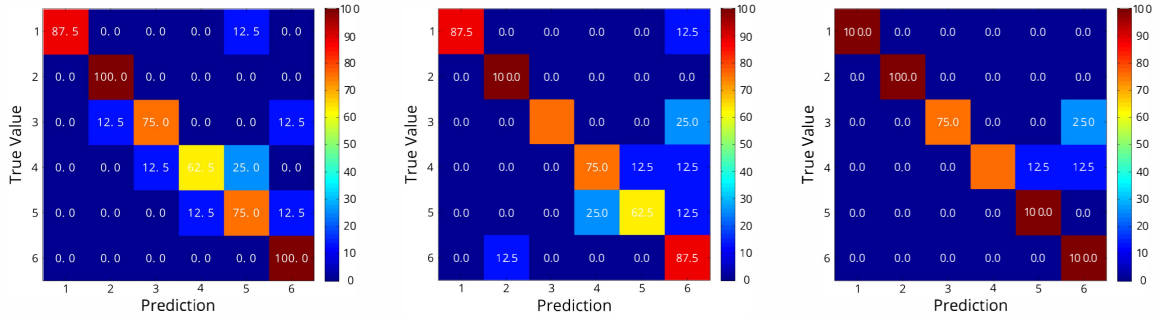
$$\begin{aligned} \min \quad & \frac{\lambda}{2}||\mathbf{W}||^2 + \frac{1}{N_1}\sum \xi_{ijk} + \frac{1}{N_2}\sum \gamma_{qmn}\xi_{qmn} \\ \text{s.t.} \quad & \mathbf{w}_{l_i}^T(\mathbf{d}_{ik} - \mathbf{d}_{ij}) \geq 1 - \xi_{ijk}\ \forall i, j, k\ \ l_i = l_j, l_i \neq l_k \\ & \mathbf{w}_p^T(\mathbf{d}_{qn} - \mathbf{d}_{qm}) \geq 1 - \xi_{qmn}\ \forall q, m, n\ \ y_q = y_m, y_q \neq y_n \\ & \mathbf{W}, \xi_{ijk}, \xi_{qmn} \geq 0 \end{aligned} \tag{1}$$

where $\mathbf{W} = [\mathbf{w}_1, \ldots, \mathbf{w}_{K_t}]$ and γ_{qmn} is defined as follows:

$$\gamma_{qmn} = \begin{cases} 1 & \text{if } \min_{z \in \mathcal{C}^t} \mathcal{M}(y_q, z) \leq \theta \wedge \\ & \min_{a \in \mathcal{C}^t, z \neq a} \mathcal{M}(y_n, a) \leq \theta \\ 0 & \text{otherwise} \end{cases} \tag{2}$$

The constraints in (1) impose that gestures of the same class should be close, while gestures of different categories should be separated by a margin of 1.

A weight vector $\mathbf{W}$ that solves (1) must satisfy two sets of constraints: one pertaining to the target data, the other to the source data. The first set contains every possible constraint we can construct on the target data, while for the second only a subset of the source classes is considered. In a nutshell, we only consider a source class if it is similar to a target class in the MMD sense. The parameter θ in (2) controls the amount of information that is transferred from the source by imposing a threshold: if the MMD between two classes is lower than θ then they are considered similar. The intuition behind the proposed approach is illustrated in Fig. 2. To solve the optimization problem in (1) we use an online learning method [19] which adopts an efficient iterative algorithm based on a stochastic gradient descent approach. The resulting algorithm is reported in Algorithm 1. Once the optimal distance vector $\mathbf{W}$ is learned, a novel test sample $\mathbf{h}$ is classified computing the $\arg\min_{l_i} \mathbf{w}_{l_i}^T \mathbf{d}(\mathbf{h}, \mathbf{h}_i)$ where $(\mathbf{h}_i, l_i) \in \mathcal{T}$.



(a) `nearest`, avg: 83.33% (b) `no-transfer`, avg: 81.25% (c) `transfer`, avg: 91.67%

Fig. 3. Confusion matrices for the three classifiers when applied to gestures 1 to 6 as performed by user 3. The classifiers are trained using two samples per gesture and tested on the others.

Gestures	Tr. Users	N	NT	T
1-6	3	84.44%	86.67%	91.11%
	2	79.17%	81.67%	86.67%
7-11	3	90.67%	97.33%	97.50%
	2	89.00%	98.50%	99.50%

Table 1. Classification accuracies obtained when classifiers are trained with data from a subset of users and tested on the rest.

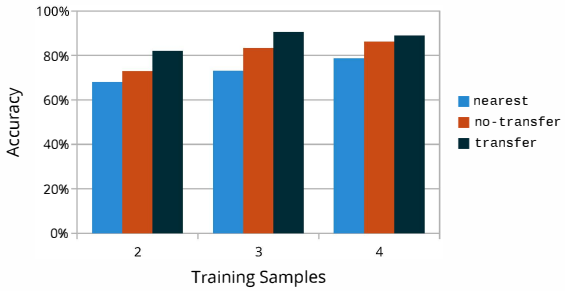


Fig. 4. Classification accuracies on gestures 1-6. Each user is considered independently and the averages are shown.

5. RESULTS

We evaluate the performance of the proposed algorithm on the set of 19 gestures shown in Figure 5. This set is split into two subsets: the source gestures, denoted with letters, and the target gestures, denoted with numbers. We recorded 15 users performing 15 repetitions of each gesture in the source set, for a total of 1800 samples, and 6 other users performing 10 repetitions of each gesture in the target dataset, for a total of 660 samples. The proposed algorithm, denoted as `transfer` (`T`), is compared against two baselines: a simple nearest neighbor algorithm using the euclidean distance, denoted as `nearest` (`N`), and a local metric learning algorithm which only uses target data, denoted as `no-transfer` (`NT`).

In a first set of experiments we compare the three classifiers in a user independent setup, *i.e.* using a subset of the users for training and the rest for testing. Specifically we consider the whole source data and we perform two experiments defining two target sets, one composed of gestures 1-6, the other of gestures 7-11. Table 1 shows the results obtained by selecting data corresponding to two or three users from the target dataset. It is clear that learning a local metric leads to a performance boost: for gestures 1-6 `no-transfer` achieves an accuracy of 86,67% and 81,67%, respectively using 3 and 2 users, against the 84,44% and 79,17% of `nearest`, while for gestures 7-11 an improvement of 7-8% is observed. Moreover, transferring knowledge from source data provides additional benefits. When the new classes are similar to the source ones, *e.g.* for gestures 1-6, the advantage of using `transfer` is more pronounced, while for gestures 7-11, which are quite dissimilar from the source, our approach correctly discards the majority of source information to avoid negative transfer.

In the last series of experiments we focus on the scenario of interest: allowing the user to reconfigure the system using few samples of novel gestures. Thus we compare the three approaches when only 2-4 samples per class are available in the target set. Results are shown in Fig. 4. It is clear that when the training set is small for the target domain, we can take advantage of the knowledge of the source, adapting it in the new context. Our approach is the only one that guarantees an accuracy greater than 80%. Obviously the performances increase as the training set size grows. Figure 3, shows the confusion matrices associated to a specific user. It is interesting to note how the accuracy of `no-transfer` is lower that the one obtained with the nearest neighbor approach, while `transfer` outperforms both thanks to the use of source data.

6. CONCLUSIONS

A smartwatch-based gesture recognition system that can be personalized by the user using a small set of samples has been presented. The system relies on a novel transfer metric learning algorithm to improve our previously proposed assistive application for the visually impaired by providing an effortless reconfiguration procedure. This algorithm has been proven to achieve good accuracy on a set of target gestures exploiting prior knowledge about another set of source gestures. Future work will be directed towards improving the gesture recognition system by allowing continuous system improve-

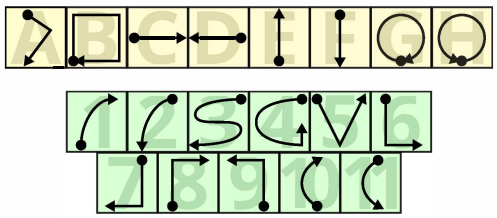


Fig. 5. The two sets of considered gestures. The source set is denoted with letters, the target set with numbers.

ment using unlabeled gestures, and towards developing additional system components *e.g.* a multimodal visual-inertial localization module [20, 21].

7. ACKNOWLEDGMENTS

This research has been partly funded by the European 7th Framework Program, under grant VENTURI (FP7-288238), and by the MIUR Smart Cities and Communities and Social Innovation program, under the grant SEAL.

REFERENCES

[1] L. Porzi, S. Messelodi, C. M. Modena, and E. Ricci, "A smart watch-based gesture recognition system for assisting people with visual impairments," in *ACM Int. Work. on Interactive Multimedia on Portable Devices*, 2013.

[2] J. Mäntyjärvi, J. Kela, P. Korpipää, and S. Kallio, "Enabling fast and effortless customisation in accelerometer based gesture interaction," in *Int. Conf. on Mobile and Ubiquitous Multimedia*, 2004.

[3] A. Akl and S. Valaee, "Accelerometer-based gesture recognition via dynamic-time warping, affinity propagation, & compressive sensing," in *IEEE Int. Conf. on Acoustics Speech and Signal Processing*, 2010.

[4] J. Liu, L. Zhong, J. Wickramasuriya, and V. Vasudevan, "uWave: Accelerometer-based personalized gesture recognition and its applications," *Pervasive and Mobile Computing*, 2009.

[5] R. Amar, S. Dow, R. Gordon, M. R. Hamid, and C. Sellers, "Mobile advice: an accessible device for visually impaired capability enhancement," in *ACM Human Factors in Computing Systems (CHI)*, 2003.

[6] F. C. Y. Li, D. Dearman, and K. N Truong, "Leveraging proprioception to make mobile phones more accessible to users with visual impairments," in *Int. ACM SIGACCESS Conf. on Computers and Accessibility*, 2010.

[7] G. Bieber, T. Kirste, and B. Urban, "Ambient interaction by smart watches," in *Int. Conf. on Pervasive Technologies Related to Assistive Environments*, 2012.

[8] E. Morganti, L. Angelini, A. Adami, D. Lalanne, L. Lorenzelli, and E. Mugellini, "A smart watch with embedded sensors to recognize objects, grasps and forearm gestures," *Procedia Engineering*, 2012.

[9] D. Bonino, F. Corno, and L. De Russis, "dwatch: A personal wrist watch for smart environments," *Procedia Computer Science*, 2012.

[10] M. Khan, S. I. Ahamed, M. Rahman, and J. Yang, "Gesthaar: An accelerometer-based gesture recognition method and its application in nui driven pervasive healthcare," in *IEEE Int. Conf. on Emerging Signal Processing Applications (ESPA)*, 2012, pp. 163–166.

[11] J. Wu, G. Pan, D. Zhang, G. Qi, and S. Li, "Gesture recognition with a 3-D accelerometer," in *Ubiquitous intelligence and computing*, 2009, pp. 25–38.

[12] Z. He, L. Jin, L. Zhen, and J. Huang, "Gesture recognition based on 3d accelerometer for cell phones interaction," in *IEEE Asia Pacific Conf on Circuits and Systems*, 2008, pp. 217–220.

[13] T. Pylvänäinen, "Accelerometer based gesture recognition using continuous HMMs," in *Pattern Recognition and Image Analysis*, 2005, pp. 639–646.

[14] S. Cho, E. Choi, W. Bang, J. Yang, J. Sohn, D. Y. Kim, Y. Lee, and S. Kim, "Two-stage recognition of raw acceleration signals for 3-d gesture-understanding cell phones," in *Int. Workshop on Frontiers in Handwriting Recognition*, 2006.

[15] S. J. Pan and Q. Yang, "A survey on transfer learning," *IEEE Trans. on Knowledge and Data Eng.*, 2010.

[16] K. Saenko, B. Kulis, M. Fritz, and T. Darrell, "Adapting visual category models to new domains," in *European Conf. on Computer Vision*, 2010.

[17] G. Costante, T. A. Ciarfuglia, P. Valigi, and E. Ricci, "A transfer learning approach for multi-cue semantic place recognition," in *IEEE/RSJ Int. Conf. on Intelligent Robots and Systems (IROS)*, 2013, pp. 2122–2129.

[18] K. M. Borgwardt, A. Gretton, M. J. Rasch, H. Kriegel, B. Schlkopf, and A. J. Smola, "Integrating structured biological data by kernel maximum mean discrepancy," *Bioinformatics*, vol. 22, no. 14, pp. e49–e57, 2006.

[19] S. Shalev-Shwartz, Y. Singer, N. Srebro, and A. Cotter, "Pegasos: Primal estimated sub-gradient solver for svm," *Mathematical programming*, vol. 127, no. 1, pp. 3–30, 2011.

[20] L. Porzi, E. Ricci, T. A. Ciarfuglia, and M. Zanin, "Visual-inertial tracking on android for augmented reality applications," in *IEEE Work. on Environmental Energy and Structural Monitoring Systems*, 2012.

[21] S. Duffner, J.-M. Odobez, and E. Ricci, "Dynamic partitioned sampling for tracking with discriminative features," in *British Machine Vision Conference*, 2009.