# EXPLOITING TRANSFER LEARNING FOR PERSONALIZED VIEW INVARIANT GESTURE RECOGNITION

*Gabriele Costante*[1] *Valerio Galieni*[1] *Yan Yan*[2] *Mario Luca Fravolini*[1] *Elisa Ricci*[1] *Paolo Valigi*[1]

[1] University of Perugia, Perugia, Italy [2] University of Trento, Trento, Italy

{paolo.valigi,mario.fravolini,elisa.ricci}@unipg.it

{gabriele.costante}@studenti.unipg.it {valeriog86}@gmail.com

{yan.yan}@disi.unitn.it

## ABSTRACT

A robust gesture recognition system is an essential component in many human-computer interaction applications. In particular, the widespread adoption of portable devices and the diffusion of autonomous systems with limited power and load capacity has increased the need of developing efficient recognition algorithms which operates on video streams recorded from low cost devices and which can cope with the challenging issue of point of view changes. A further challenge arises as different users tend to perform the same gesture with different styles and speeds. Thus a classifier trained with gestures data of certain set of users may work poorly when data from other users are being processed. However, as often a mobile device or a robot are intended to be used by a single or by a small group of people, it would be desirable to have a gesture recognition system designed specifically for these users. In this paper we introduce a novel approach to face the problems of view-invariance and user personalization in the context of gesture interaction systems. More specifically, we propose a domain adaptation framework based on a feature space augmentation approach operating on robust view-invariant Self Similarity Matrix descriptors. To prove the effectiveness of our method a dataset corresponding to 17 users performing 10 different gestures under 3 point of views is collected and an extensive experimental evaluation is performed.

***Index Terms***— Gesture recognition, low cost cameras, view invariance, transfer learning.

## 1. INTRODUCTION

In recent years, due to the widespread adoption of mobile devices and to the advent of low cost RGB-D sensors, we are witnessing a renewed interest towards developing novel gesture interaction systems. Many solutions can be devised, mostly depending on the employed sensors (*e.g.* inertial sensors, traditional webcams, RGB-D cameras) and on the application (*e.g.* human-robot cooperation, human-computer interaction). In this paper we focus our attention on low cost RGB cameras as we ultimately target the application of designing a gesture-recognition approach used for imposing commands to an autonomous system with limited power and load capacity, *e.g.* a quadcopter.

In the computer vision community, the problems of gesture and action recognition have received considerable attention in the past [1]. Many challenges arise, mainly due to the varying position of the target with respect to the camera, to the user-specific way of executing a particular gesture and, more in general, to other environmental factors (*e.g.* changes in illumination conditions).

In this paper we introduce an approach to effectively cope with the first two issues. We propose a novel gesture recognition method which takes advantage of the recent Self-Similarity Matrix (SSM) descriptors [2] to compute features which are robust to small changes of point of view and adopts an efficient transfer learning methodology to improve recognition accuracy through user personalization.

User personalization is particularly important in the considered scenario of human-robot interaction since after a quadcopter is bought, typically a single or a limited set of users will be demanded to command it in order to perform specific tasks. In general other applications can benefit from a gesture recognition system targeted to specific users, *e.g.* applications involving personal portable devices. We consider the scenario where we have only few gestures samples (the target samples) from a specific user, collected during a very short configuration phase. As a classifier trained only on these data will perfom poorly, we propose to exploit previous knowledge available in form of gesture sequences (the source data) collected by other users (depending on the application and on the specific set of gestures considered, these data can be gathered easily by downloading them from the web). Transfer learning operates by transferring useful information from source data in order to improve classification perfomance on target data.

We evaluate our approach on a dataset consisting of 10 gestures (corresponding to the user drawing in the air the ten digits 0-9) performed by 17 targets. Each gesture is repeated three times. Each repetition correspond to a different posi-

978-1-4799-2893-4/14/$31.00 ©2014 IEEE

tion of the target with respect to the camera. Many previous works have addressed the gesture recognition problem [3, 4, 5], however, to our knowledge, there are no view invariant approaches performing domain adaptation on specific users.

## 2. RELATED WORKS

Gesture recognition is an active research area in computer vision [1]. In general visual based gesture interaction systems suffer from lack of robustness due to point of view changes, intra-class variations and self-occlusions. To overcome the issue of point of view variations many works have focused on transferring visual features across views [6, 7, 8, 9, 10] or using view-invariant features [2, 11].

Bag-of-words models are probably the most widely used approaches to address the problem of gesture recognition [12, 13, 14]. Visual vocabularies are learned from clustering spatio-temporal feature descriptors, then the video features are computed by projecting the associated descriptors into the precomputed vocabulary. Another class of methods exploits autoregressive (AR) models [15, 16]. In [17] a Tensor Canonical Correlation Algorithm (TCCA) is proposed to compute the similarity among actions while in [18] SIFT features in combination with CCA are used to perform gesture recognition. In this paper we adopt the view invariant descriptors proposed in [2], evaluating their suitability to our scenario and to the chosen set of gestures. In fact, in [2] a different problem is considered, *i.e.* action recognition in a multi-camera setup.

Recently, due to the need of deploying systems able to reconfigure themselves according to changes of environmental conditions, transfer learning approaches have become popular in the computer vision community. A good survey on these techniques is presented in [19]. Zheng *et al.* [20] propose to perform knowledge transfer from source data to target data using the Maximum Mean Discrepancy criterion to evaluate how much to transfer. In [21] transfer learning is adopted to improve visual object categorization performances, adapting classifiers previously learned to a novel scenario. In [22] a risk based transfer learning framework is proposed for place recognition. In this work we consider the adaptation framework proposed in [23] as it is simple, easy to implement and does not require additional computation, *e.g.* the MMD distance. No previous works have exploited the feasibility of the method in [23] in the context of gesture recognition applications.

## 3. USER INDEPENDENT GESTURE RECOGNITION

In this section we first introduce the Self Similarity Matrix (SSM) features [2], then we present the adopted transfer learning approach.

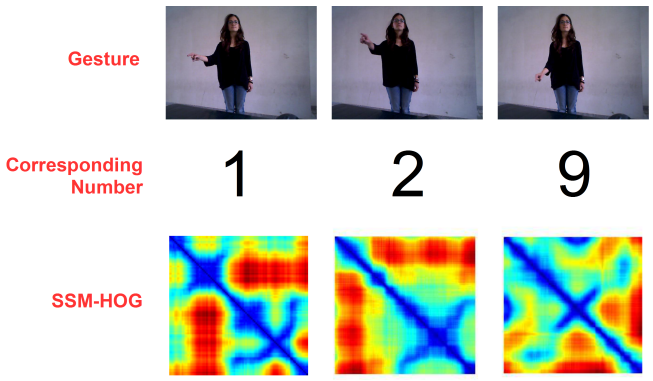


**Fig. 1**: An example of SSM features extracted from three sample videos corresponding to the digits 1, 2, 9. In the first row frames extracted from the video sequences are shown. The second row depicts the corresponding digits and the third the relative SSM matrices.

### 3.1. Self Similarity Matrix Descriptors

In a recent work Junejo *et al.* [2] proved that SSM descriptors are very robust features for view invariant action recognition. Given a video sequence composed by $N$ frames $\{\mathcal{F}_1, \mathcal{F}_2, \ldots, \mathcal{F}_N\}$, a SSM is defined as:

$$[s_{ij}]_{i,j=1,2,\ldots,N} = \begin{bmatrix} 0 & s_{12} & s_{13} & \ldots & s_{1N} \\ s_{21} & 0 & s_{23} & \ldots & s_{2N} \\ s_{31} & s_{32} & 0 & \ldots & s_{3N} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ s_{N1} & s_{N2} & s_{N3} & \ldots & 0 \end{bmatrix} \quad (1)$$

where $s_{ij}$ is an appropriate distance functions between frames $\mathcal{F}_i$ and $\mathcal{F}_j$. In this work we compute a Histogram of Oriented Gradients (HOG) descriptor [24] for each frame and use the euclidean distance to calculate $s_{ij}$. Clearly the elements on the diagonal of the SSM are zero since they correspond to a frame compared to itself. The final features associated to a video sequence are obtained by computing for each point on the diagonal of the SSM three descriptors corresponding to three different diameters extracted from the log-polar domain (28, 42 and 56 frames per diameter) and then using a bag-of-word approach (we used a codebook of 500 words) [2]. Figure 1 shows an example of the SSM matrices computed on three sample video sequences.

### 3.2. Domain Adaptation

In this paper we propose to build a user-specific gesture recognition system by adopting the domain adaptation method proposed in [23]. Domain adaptation approaches operate by defining a source and a target data set. In our application the *source* domain consists of a set of gestures acquired from several subjects. Considering a new user corresponding to the

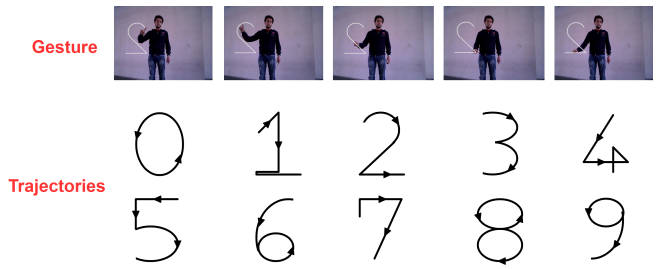


**Fig. 2**: Frames extracted from a sequence depicting a user performing the gesture '2' (top). The considered gesture categories, *i.e.* the 0-9 digits (bottom).

*target* domain, it would be desirable to achieve high recognition accuracy by avoiding to collect data of the user itself or allowing a very short reconfiguration phase. Indeed, if the novel user performs gestures in a different scenario, with different style and speed with respect to the training data, a traditional gesture recognition approach performs poorly. Thus, a desirable property of a recognition system should be the ability to reconfigure itself easily using few sequences from the novel user, *e.g.* 2-3 samples. More rigorously, we are dealing with a supervised domain adaptation problem, where a great amount of labeled data from the source domain is available together with a small portion of labeled samples form the target domain. Formally we define a source set $\mathcal{S} = \{(\mathbf{v}_1^s, y_1^s), (\mathbf{v}_2^s, y_2^s), \ldots, (\mathbf{v}_P^s, y_P^s)\}$, and a target set $\mathcal{T} = \{(\mathbf{v}_1^t, y_1^t), (\mathbf{v}_2^t, y_2^t), \ldots, (\mathbf{v}_M^t, y_M^t)\}$, where $\mathbf{v}_i \in I\!R^d$ and $y_i \in I\!R$ are respectively the SSM descriptor computed as discussed in the previous subsection and the associated label. $P$ and $M$, $M << P$, are the dimension of the source and the target set and $y_i^t$, $y_i^s \in \{1, 2, \ldots, K\}$ where $K$ is the number of different gesture categories. To pursue domain adaptation we consider the feature augmentation approach proposed in [23]. We define two mappings $\psi^s$ and $\psi^t$ to be applied respectively to source and target data:

$$\begin{aligned} \psi^s(\mathbf{v}_i^s) &= \langle \mathbf{v}_i^s, \mathbf{v}_i^s, \mathbf{0} \rangle \\ \psi^t(\mathbf{v}_i^t) &= \langle \mathbf{v}_i^t, \mathbf{0}, \mathbf{v}_i^t \rangle \end{aligned} \tag{2}$$

and consider the associated augmented sets $\mathcal{S}'$, $\mathcal{T}'$. It is evident that the final feature space is three times larger than the original one. Based on the constructed augmented training set a standard SVM classifier is trained using both $\mathcal{S}'$ and $\mathcal{T}'$. The resulting vector $\mathbf{w} \in I\!R^{3d}$, corresponding to the optimal separating hyperplane, is made of three parts: the first $d$ components correspond to features common to both source and target data, while the remaining $2d$ components are associated to source specific and target specific features. Therefore, when training a SVM not only the optimal hyperplane for gesture recognition accuracy is obtained, but also the relative importance of each feature is determined. As demonstrated in the experimental section, through this approach it is possible to effectively transfer knowledge from source data.

## 4. RESULTS

In this section we first introduce the experimental setup, then we discuss the results of our evaluation.

### 4.1. Experimental setup

In order to prove the effectiveness of our approach we recorded several videos depicting 17 targets that perform a set of gestures (specifically gestures corresponding to the numbers 0-9). Each gesture is executed three times under different view points: in the first repetition the user is exactly in front of the camera (*center*) while in the second and in the third the position is slightly turned and shifted respectively on the *left* and *right* side of the camera field-of-view. The sequences have been recorded with a standard webcam (resolution 640 $\times$ 480 pxls). Figure 2 depicts the ten considered gesture categories together with some frames extracted from one video sequence.

### 4.2. Quantitative Evaluation

We perform two series of experiments: the first aims to assess the robustness of our approach with respect to changes of point of view, the second evaluates the effectiveness of the proposed transfer learning solution.

To prove the suitability of SSM-HOG features in our scenario we perform a set of experiments using a leave-one-user-out protocol. We pick a subset of 6 users and randomly select the sequences of 5 users as training set and the video associated to the remaining one as test data. We consider a standard SVM classifier with RBF kernel and perform two different experiments: in the first for each gesture we select data corresponding to five users and a single view (*central, left, right*) as training set and employ sequences of the same view as test data, while in the second test the other two views are used for testing. The optimal regularization parameter and the best $\sigma$ value for the RBF kernel are set with cross-validation. As a baseline we also consider a standard nearest neighbor (NN) classifier. Results are shown in Table 1. The average accuracy of three repetitions, corresponding to different views, are reported. Two interesting observations can be made from these tests. Firstly, the SVM clearly outperforms the nearest neighbor classifier. Secondly, SSM feature guarantees view invariance since the tests with different views lead to similar performances to the ones corresponding to the same view. Moreover high variations in accuracy are observed among different tests. This is caused by the strong variability between gesture executions of different users.

In a second series of experiments we evaluate the proposed domain adaptation approach. We divide the entire dataset into a *target* set containing data from a single user and a single point of view and a *source* set with data of the remaining users. The rest of the sequences from the selected

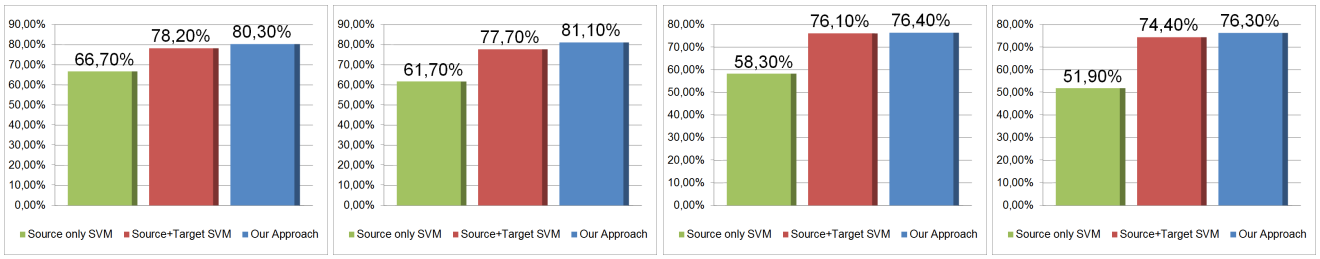


**Fig. 3**: Results of the domain adaptation experiments. Each bar graph corresponds to a six-users test. The user indexes in each group from left to right are respectively (1 4 2 12 17 11), (3 10 13 5 9 16), (6 15 14 7 8 10) and (3 1 8 12 16 10).

**Table 1**: View invariance tests: comparison of different methods (accuracy %).

| | User 1 | User 2 | User 3 | User 4 | User 5 | User 6 | Average |
|---|---|---|---|---|---|---|---|
| NN (same view) | 81.2 | 70.4 | 42.5 | 41.1 | 53.3 | 49.7 | 56.3 |
| NN (different views) | 80.3 | 68.4 | 43.0 | 41.3 | 52.8 | 50.1 | 55.9 |
| SVM (same view) | 90.0 | 83.3 | 50.0 | 50.0 | 60.0 | 64.3 | 66.1 |
| SVM (different views) | 86.7 | 88.3 | 56.7 | 56.7 | 46.7 | 65.0 | 66.7 |

user are taken as test set. We perform four experiments, corresponding to four subgroups of six randomly chosen users. Results are presented in Fig.3. Performances are averaged with respect to all the point of views and all the possible divisions of source and target data (*i.e.* taking all the six users to construct the target set). We compare our view-invariant transfer learning approach against two baselines: a *SVM - source only*, *i.e.* a SVM trained using source data, and a *SVM - source + target*, *i.e.* a standard SVM trained with sequences from the source and sequences corresponding to one view of the target user. From Fig.3 it is evident that the *SVM - source only* always achieves poor performances with respect to the other two methods. Instead a great boost in terms of accuracy is achieved in case of the *SVM - source + target*. Clearly in this case the algorithm takes advantage of user specific information to build a more accurate gesture classification model. While the *SVM - source only* gets only 66.7%, 61.7%, 58.3% and 51.9% in the four six-users tests, the *SVM - source + target* improves the accuracy by at least 10% on average in each test. Finally, our approach outperforms the *SVM - source + target*. The feature augmentation procedure allows the SVM to effectively "understand" which parts of the source data are common to the target sequence and which are domain-specific.

In the last series of experiments we evaluate the classification performances varying the number of users in the source set. In particular we compare the *same view* and the *different views* setup using both the nearest neighbor (NN) and the SVM classifiers against our transfer learning method (Fig. 4). The results confirm the trends observed in the previous experiments: the nearest neighbor always performs worse than the SVM, while our approach outperforms the baselines. Only the *SVM - source + target* reaches performances similar to our method, but we always achieve 3-4 % more.

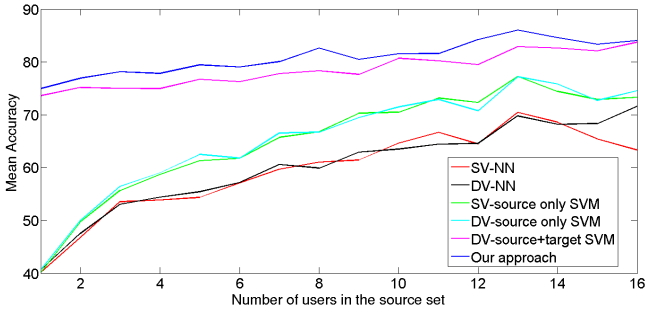


**Fig. 4**: Performances varying the number of training users. *SV* and *DV* refer to the tests considering respectively the same view and the different views while *NN* indicates the nearest neighbor approach.

## 5. CONCLUSIONS

In this paper we consider the problem of robust gesture recognition from video streams captured by low cost devices and we specifically address two important challenges: (i) view-invariance and (ii) user customization. We propose an approach which adopts powerful SSM descriptors to cope with issues related to variations of point of view, and on a effective domain adaptation algorithm based on feature augmentation, to improve recognition accuracy when the system is intended to be used by a specific user. Future works include the integration of motion features into the SSM framework to further improve performances. It would be also interesting to exploit the feasibility of our method in case of video streams gathered from RGB-D sensors and to integrate our approach into an autonomous quadcopter for human-robot interaction applications.

## 6. REFERENCES

[1] Jose M. Chaquet, Enrique J. Carmona, and Antonio Fernández-Caballero, "A survey of video datasets for human action and activity recognition," *Computer Vision Image Understanding*, vol. 117, no. 6, pp. 633–659, 2013.

[2] Imran N. Junejo, Emilie Dexter, Ivan Laptev, and

Patrick Perez, "View-independent action recognition from temporal self-similarities," *Transaction on Pattern Analysis and Machine Intelligence*, vol. 33, no. 1, pp. 172–185, 2011.

[3] L. Ballan, M. Bertini, A. Del Bimbo, L. Seidenari, and G. Serra, "Effective codebooks for human action representation and classification in unconstrained videos," *Transaction on Multimedia*, vol. 14, no. 4, pp. 1234–1245, 2012.

[4] S.M.A. Hussain and A.B.M.H. Rashid, "User independent hand gesture recognition by accelerated DTW," in *International Conference on Informatics, Electronics Vision (ICIEV)*, 2012, pp. 1033–1037.

[5] Jiayang Liu, Lin Zhong, Jehan Wickramasuriya, and Venu Vasudevan, "uWave: Accelerometer-based personalized gesture recognition and its applications," *Pervasive and Mobile Computing*, vol. 5, no. 6, pp. 657–675, 2009.

[6] Ali Farhadi and Mostafa Kamali Tabrizi, "Learning to recognize activities from the wrong view point," in *European Conference on Computer Vision (ECCV): Part I*, 2008, pp. 154–166.

[7] Ruonan Li, "Discriminative virtual views for cross-view action recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2855–2862.

[8] Jingen Liu, M. Shah, B. Kuipers, and S. Savarese, "Cross-view action recognition via view knowledge transfer," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3209–3216.

[9] Anoop K. Rajagopal, Ramanathan Subramanian, Radu L. Vieriu, Elisa Ricci, Oswald Lanz, Kalpathi Ramakrishnan, and Nicu Sebe, "An adaptation framework for head-pose classification in dynamic multi-view scenarios," in *Asian conference on Computer Vision (ACCV) - Volume Part II*, 2013, pp. 652–666.

[10] Yan Yan, Gaowen Liu, Elisa Ricci, and Nicu Sebe, "Multi-task linear discriminant analysis for multi-view action recognition," in *International Conference on Image Processing (ICIP)*, 2013, pp. 2842–2846.

[11] Cen Rao Alper Yilmaz and Mubarak Shah, "View-invariant representation and recognition of actions.," *International Journal of Computer Vision*, vol. 50, no. 2, pp. 203–226, 2002.

[12] Heng Wang, Muhammad Muneeb Ullah, Alexander Kläser, Ivan Laptev, and Cordelia Schmid, "Evaluation of local spatio-temporal features for action recognition," in *British Machine Vision Conference (BMVC)*, 2009, p. 127.

[13] P. Dollar, V. Rabaud, G. Cottrell, and S. Belongie, "Behavior recognition via sparse spatio-temporal features," in *International Conference on Computer Communications and Networks (ICCCN)*, 2005, pp. 65–72.

[14] Piotr Bilinski and Francois Bremond, "Evaluation of local descriptors for action recognition in videos," in *International Conference on Computer vision systems (ICVS)*, 2011, pp. 61–70.

[15] Payam Saisan, Gianfranco Doretto, Ying Nian Wu, and Stefano Soatto, "Dynamic texture recognition," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 58–63.

[16] A. Bissacco, A. Chiuso, Yi Ma, and S. Soatto, "Recognition of human gaits," in *Computer Vision and Pattern Recognition (CVPR)*, 2001, vol. 2, pp. 52–57.

[17] Tae-Kyun Kim and R. Cipolla, "Canonical correlation analysis of video volume tensors for action categorization and detection," *Transaction on Pattern Analysis and Machine Intelligence*, vol. 31, no. 8, pp. 1415–1428, 2009.

[18] Tae-Kyun Kim and Roberto Cipolla, "Gesture recognition under small sample size," in *Asian Conference on Computer vision (ACCV) - Part I*, 2007, pp. 335–344.

[19] Sinno Jialin Pan and Qiang Yang, "A survey on transfer learning," *Transaction on Knowledge and Data Engineering*, vol. 22, no. 10, pp. 1345–1359, 2010.

[20] Vincent Wenchen Zheng, Derek Hao Hu, and Qiang Yang, "Cross-domain activity recognition," in *International conference on Ubiquitous computing (Ubicomp)*, 2009, pp. 61–70.

[21] Kate Saenko, Brian Kulis, Mario Fritz, and Trevor Darrell, "Adapting visual category models to new domains," in *European conference on Computer vision (ECCV): Part IV*, 2010, pp. 213–226.

[22] Gabriele Costante, Thomas A. Ciarfuglia, Paolo Valigi, and Elisa Ricci, "A transfer learning approach for multi-cue semantic place recognition," in *International Conference on Intelligent Robots and Systems (IROS)*, 2013, pp. 2122–2129.

[23] Hal Daumé, III, Abhishek Kumar, and Avishek Saha, "Frustratingly easy semi-supervised domain adaptation," in *Workshop on Domain Adaptation for Natural Language Processing (DANLP)*, 2010, pp. 53–59.

[24] Navneet Dalal and Bill Triggs, "Histograms of oriented gradients for human detection," in *Conference on Computer Vision and Pattern Recognition (CVPR)*, 2005, pp. 886–893.