LEARNING SYMMETRIC FACE POSE MODELS ONLINE USING LOCALLY WEIGHTED PROJECTRON REGRESSION

Jawad Nagi, Gianni A. Di Caro, Alessandro Giusti, Luca M. Gambardella

Dalle Molle Instituite for Artificial Intelligence (IDSIA), Lugano, Switzerland
{jawad,gianni,alessandrog,luca}@idsia.ch

ABSTRACT

Human localization is fundamental in human centered computing and human-robot interaction (HRI), as human operators should be localized by robots before being actively serviced. This paper proposes a simple and efficient approach for estimating the distance and orientation of an human, from a single robot-acquired image. We adopt a simple combination of multiple Haar feature-based classifiers to compute *face scores*, that represent the probability that the detected face is acquired from each of a predefined set of poses. Using the Locally Weighted Projectron Regression (LWPR), an online incremental regression-based learning scheme, we can reliably learn and predict the pose of a human face in real-time at a low computational cost. The accuracy, robustness, and scalability of the obtained solutions have been verified through emulation experiments performed on a large data set of real data acquired by a networked swarm of robots.

Index Terms— Head pose estimation, multi-camera, face scores, online incremental learning, non-linear regression

1. INTRODUCTION

Human localization in smart indoor environments is important in many applications, such as visual surveillance, monitoring and ambient assisted-based living. Face pose estimation has been an active topic in the computer vision community, due to its significant role in many real-world applications such as: gaze detection, multi-view face recognition and human-robot localization. It is a challenging problem due to factors associated with illumination conditions, facial expressions, subject variability and camera distortion.

Being aware of a human's location (i.e., localization) is a precondition for human-centered computing applications (e.g. human-robot interaction (HRI)). There are several ways of localization, including vision-based methods and RFID-based methods. Vision-based methods have recently gained more attention due to their advantages of requiring no additional wearable sensing devices. In this work, we focus attention towards a vision-based method for human and multi-UAV localization, which has no restrictions on a human's movement or posture, and is robust in real situations.

Over the years, many techniques have been proposed for face pose estimation from a monocular camera. They can be categorized in three different classes: model-based, appearance-based and hybrid approaches. Model-based approaches typically rely on the use of geometrical properties, such specific sets of facial features such as eyes, nose, mouth are used for face pose estimation [1, 2]. On the other hand, appearance-based approaches use the entire face (head) to model and learn from training data [3] and formulate the face pose estimation problem as a supervised machine learning task. Appearance-based works have generally made use of classifiers such as Support Vector Machines (SVMs) [4] and regression-based techniques such as Support Vector Regression (SVR) [5, 6, 7].

Hybrid approaches are a combination of model-based and appearance-based approaches [8, 9, 10]. Each category of approaches has its limits and constraints. Hybrid approaches generally provide better performance, but are computationally expensive and not suitable for real-time implementation, and although model-based methods are fast and simple, they are sensitive to occlusion and usually require high resolution images which may be not available in many applications such as driver monitoring or video surveillance. Generally in appearance-based approaches, facial images are compared with a set of facial appearance templates to find the most similar match, however the computational cost of comparing each image with a large number of templates is computationally expensive. To overcome these limitations, we direct our attention towards regression-based approaches, as they can easily assign a discrete pose to a set of computed facial features.

Non-linear regression-based methods have demonstrated to be effective for face pose estimation tasks [11, 12, 13, 14, 6]. In [14], gradient features were computed from face images and then fed into a SVR. A kernalized version of the Partial Least Squares (PLS) or (Ridge Regression (RR)) was adopted in [12], which notably improved the performance of face pose estimation. One difficulty faced by these regression-based approaches was determining an appropriate kernel space for mapping facial features. In a previous task [15], we adopted the RR approach for online multi-view point learning of hand gestures (sensed by a swarm of foot-bot robots [16]), however this task was modeled as a multi-class classification problem.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

Despite the above mentioned methods, we consider that, given a face image, face detectors can assign a confidence score to a detection. If the detection threshold is set low, this results in a number of sub-windows being clustered around the face. We use this approach to determine face quality measures. For online learning and prediction of the face pose, we augment the face quality measures with the Locally Weighted Projectron Regression (LWPR), an online regression approach, that uses a mixture of locally linear kernalized regressors. This paper is organized as follows. Section 2 presents the contributions of our work, i.e., the algorithms and methods for solving the aforementioned issues. Section 3 presents the findings of the experiments evaluated on real data, and Section 4 presents concluding remarks.

2. IMPLEMENTATION

2.1. Face Detection and Tracking

Face detection allows a swarm of UAVs to identify position and visual orientation of human operators with respect to their location. In turn, face detection is functional to determine the relative angular, radial, and altitude position of UAVs with respect to a human. We adopt the notion of face detection to create a normalized and user-centric view of the human, from the point of view of multiple UAVs. At first, the task of each robot in our networked swarm of UAV drones is to detect a human for interaction using its onboard camera. To achieve this, we use the front-mounted cameras of the Parrot A.R. Drone 2.0 quadcopters, that acquire images in a native HD resolution of 1280×720 pixels at 30 fps. Face detection is performed using the OpenCV implementation of the Viola-Jones face detector [17]. As face detectors are insensitive to small changes in scale or position, multiple face detection windows are often clustered around a face, as illustrated in Figure 1.

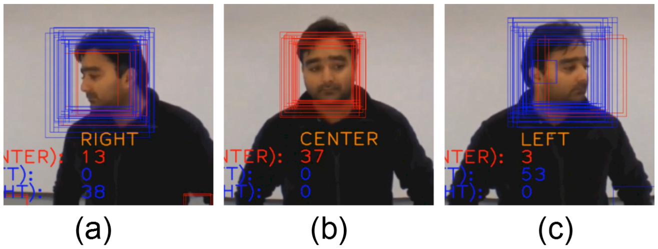

Fig. 1: Face pose estimation on an airbone UAV. Identified face poses: (a) right, (b) center, (c) left.

We use the notion of the *detected sub-windows* (i.e., output of the Haar face detector), as a measure to assess the quality of the detected face. These sub-windows represent the confidence of a face detection classifier. The larger the number of sub-windows detected from a face, the higher confidence the classifier has in detecting the face, and vice versa. In practice, using the OpenCV face detector, we set the parameter specifying the "number of neighbours each candidate sub-window should retain" to be maximum, which identifies all groups of *neighbouring sub-windows* clustered around the face. Furthermore, as rapid ego-motion of airborne camera's onboard UAVs can lose a detected face or detect false positives, we adopt a Kalman Filter for tracking a detected face. In order to determine which candidate face sub-window to use as an input for tracking, a nearest neighbour strategy is employed using the Mahalanobis distance computed from the covariance of the detected face sub-windows.

2.2. Face Pose Estimation

Inspired by the well known AdaBoost technique [18] that implements a robust face detector capable of detecting not only frontal faces, but lateral (left and right) profiles as well, in this work we propose a combination of multiple Haar face detectors that can identify a human face from a 3-dimensional perspective. We use the *number of neighbouring sub-windows* detected around the face to estimate the pose (position) of a human's face (in terms of a meaningful score) from a robot's point of view.

In our approach, we employ two pre-trained Haar face detectors, one classifier FC_f trained on the 'frontal views' of the face profile, while the other classifier, FC_s on the 'side profiles' of the face, as illustrated in Figure 1, where the red-coloured sub-windows show face detections from FC_f and the blue-coloured sub-windows show detections from FC_s. For every acquired image frame I, four relative *face quality measures* $Fm = \{Fm_f, Fm_{ff}, Fm_s, Fm_{sf}\}$ are computed using FC_f and FC_s. Each face measure in Fm represents the *number of the neighbouring sub-windows detected around the face* from different poses (i.e., front, side, and flipped):

1. FC_f is run on I to obtain Fm_f;

2. I is flipped horizontally (180°; horizontal shift) to obtain I_h. I_h is then processed by FC_s to produce Fm_{ff};

3. FC_s is run on I to get Fm_s;

4. FC_s is run on I_h (obtained in Step 2) to obtain Fm_{sf}.

Using the these face quality measures, three (3) *face scores* $\{S^c, S^r, S^l\}$ are derived for representing the current face pose from a robot's point of view, where $S^c = Fm_f + Fm_{ff}$, $S^r = Fm_s$, and $S^l = Fm_{sf}$. Large values of a face score mean that a face is detected with a high confidence, and vice versa for low scores. In simpler words, when a UAV is positioned directly in front of a human (frontal view of the face), the value of S^c is higher than S^r and S^l. If the UAV is positioned towards the left or right side of the human (side profile), then the value of S^l or S^r respectively will be higher than all other scores. If all three scores are below the threshold S_{TH}, then the human is not present in the robot's field of view, or is too far away to be detected.

Face pose estimation also needs to take into account the relative distance between a human and a robot, which provides another reliable measure to aid human and multi-robot localization. The distance between a robot and a human is estimated by using the face quality measures in Fm, and calculating the total sum of the *area of all detected sub-windows*, $d = (\sum_T^i {F_A(i)}/{T})$, in $F_A = \{Fm_{area}(i), ..., Fm_{area}(T)\}$, where $T = Fm_f + Fm_{ff} + Fm_s + Fm_{sf}$. Large values of d represent that the robot is near to the human, whereas smaller values indicate the robot is far from the human. We use $\{S^c, S^r, S^l\}$ and d (computed from every image), as features for learning and estimating the face pose.

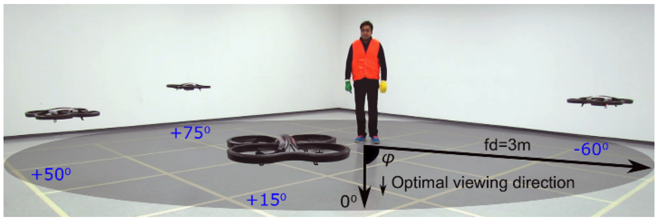

Fig. 2: Spatial arrangements of UAVs for human-UAV localization and data acquisition.

2.3. Dataset Acquisition

We build dataset of images using a swarm of four (4) Parrot drones, by acquiring face pose images in a resolution of 1280×720 pixels. Using 4 drones we could acquire a relatively large amount of images of a human's face from multiple points of view. To acquire the dataset, the UAVs are positioned at around the human using the multi-robot formation illustrated in Figure 2. Using this configuration, each robot acquired and stored approximately 1200 unprocessed images with known ground truth information (θ, D), while a human operator rotated his face in a semi-circular plane of $[0, 180°]$. The process is repeated 5 times, once for a different distance $D = \{1, 2, ..., 5\}$m between the UAVs and the human, which results in a dataset of approximately $24,000$ images, acquired by the swarm from $4 \times 5 = 20$ different viewpoints.

2.4. Locally Weighted Projection Regression

The Locally Weighted Projection Regression (LWPR) is a family of online incremental learning algorithms that performs piecewise linear function approximation using regression. By detecting locally redundant or irrelevant input dimensions, LWPR locally reduces the dimensionality of the input data by finding local projections using Partial Least Squares (PLS) regression [19].

In this work, we employ LWPR to learn a non-linear regression function from training data (by means of piecewise linear models called *receptive fields*) that incrementally arrive as input-output tuples ($\mathbf{x}_i$,y_i), considering multi-variate output data. Learning online and incrementally (as data arrives) involves automatically determining the appropriate number of receptive fields (i.e., local models) [20]. In supervised learning algorithms, if $\mathbf{x}_i$ denotes a set of *features* computed from a single image, then y_i represents its respective *target label*. Thus, the LWPR regression function can be constructed by blending local linear models $\Psi_k(\mathbf{x})$ in the form [21]:

$$f(\mathbf{x}) = \frac{1}{T(\mathbf{x})} \sum_{k=1}^{K} w_k \Psi_k(\mathbf{x}), \qquad T(\mathbf{x}) = \sum_{k=1}^{K} w_k(\mathbf{x}) \qquad (1)$$

where $T(\mathbf{x})$ represents the normalization factor and $w_k(\mathbf{x})$ is a *locality kernel* (i.e., the activation of a receptive field) that defines the area of validity of the local models (receptive fields), which we model as a *Gaussian* (RBF) function, in order to fit the data using non-linear regression [21]:

$$w_k(\mathbf{x}) = \exp\left(-\frac{1}{2}(\mathbf{x} - \mathbf{c}_k)^T \, \mathbf{D}_k(\mathbf{x} - \mathbf{c}_k)\right) \qquad (2)$$

where $\mathbf{c}_k$ is the centre of the k^{th} linear model (or receptive field), $\mathbf{D}_k$ is its positive semi-definite distance matrix that determines the size and shape of the neighbourhood contributing to the local model.

For learning the linear models $\Psi_k(\mathbf{x})$, in this work we employ an online formulation of weighted PLS regression within each local model of LWPR to fit the hyperplane. Given a query point $\mathbf{x}$, each linear model calculates a prediction $\hat{y}_k(\mathbf{x})$. The output of the LWPR is the *normalized weighted mean* (i.e., linear combination) of all K linear models represented by, $\hat{y} = {(\sum_{k=1}^{K} w_k \hat{y}_k)}/{(\sum_{k=1}^{K} w_k)}$ [22]. As a significant computational advantage, we expect that far fewer projections than the actual number of input dimensions are needed for accurate learning.

3. EXPERIMENTAL RESULTS

To demonstrate the capabilities of the developed system, we performed experiments investigating the performance, robustness, and efficiency of the solution proposed in Section 2. The dataset described above has been used for running quantitative *emulation* experiments: face observations are sampled from this dataset of real images for learning and prediction.

We use subsets of images from the dataset to train and validate the LWPR. For training (learning) and testing (validation) we use $\mathbf{x}_i = \{S_i^c, S_i^r, S_i^l, d_i\}$ to be the four (4) *facial features* and $y_i = (\theta, D)$ their respective *target labels*. Using a Gaussian (non-linear) kernel, the LWPR maps these features into a *face pose* ϕ, that we project onto a horizontal plane $[0, 180°]$ (with d serving as a normalization factor). Thus, an ordered pair (ϕ, d) computed from a single face image represents the *angular distance* between a human and the UAV, that is a useful measure to aid human-swarm localization.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE

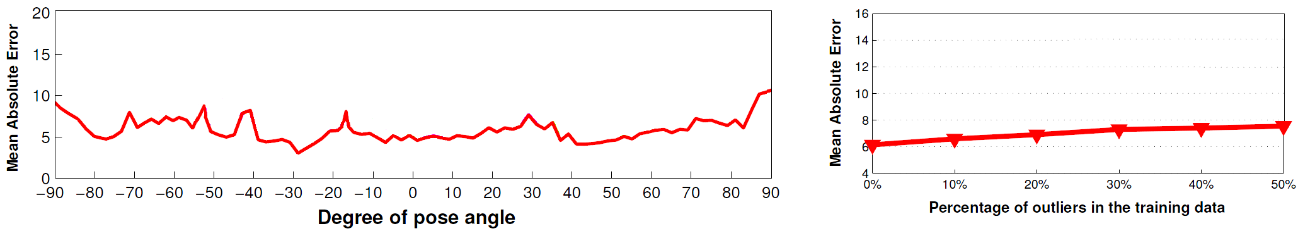

Fig. 3: MAE of face pose detection for horizontal plane $[0, 180°]$.

We study the face pose detection accuracy of a single robot as a function of its angular distance (r_ϕ, r_d). Using subset of images from the dataset, we estimate (r_ϕ, r_d) from every image and compute the *average pose accuracy* P_{acc} using the ground truth data (i.e., difference between the actual and predicted angular distances), as reported in Figure 4. Robots positioned at distances between $d = [1, ..3]$m in the central locations provide good recognition accuracies (up to 97%). With the increase of the radial distance between the human and the robots (e.g. $d \geq 3$m), face detection performance systematically degrades, since the face is not detected reliably. As a result, we consider $(2 \leq d \leq 3\text{m})$ to be a reasonably safe proximity for humans to interact with airborne UAVs.

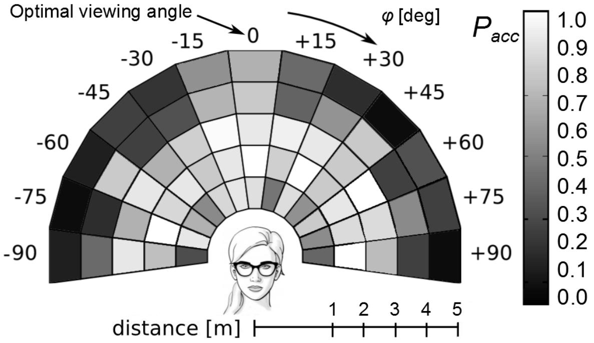

Fig. 4: Face pose detection accuracy of a single robot as a function of angular distance (ϕ, d).

Averaging the difference between the ground truth data and the predicted face poses, we can compute the Mean Absolute Error (MAE), as illustrated in Figure 3 (left). This indicates that our approach is robust to a variety of poses in a $[0, 180°]$ plane, with an average MAE of less than 10, which is stable with pose variations in the range $[-80, 80°]$. Furthermore, we intentionally vary the percentage of outliers in all training data from 0% (uncorrupted) to 50%, as depicted in Figure 3 (right). As the number of noisy samples increase in the training set, the MAE increases higher. This shows that our approach works robustly with noisy (corrupted) data, which is common in real-world applications.

Lastly, a comparative analysis of the LWPR approach (online) with other regression-based learning schemes used in context of face pose estimation, namely, SVR (batch) and RR (online) schemes is performed, as illustrated in Figure 5. It is observable that LWPR requires a smaller amount of samples to provide a good recognition accuracy as compared to RR. However, SVR has a faster convergence rate as uses one-shot batch-learning and can generalize better with a smaller number of training samples, as compared to online learning algorithms. Thus, it is significant to say that LWPR works best amongst other online regression approaches due to sparsity regularization in the model, and is robust to outliers.

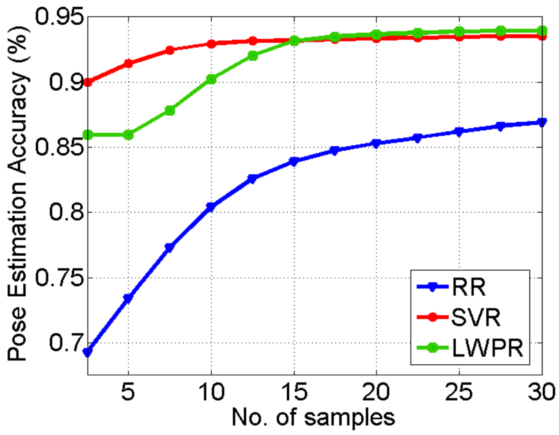

Fig. 5: Performance comparison of different regression-based learning schemes.

4. CONCLUSION

In this paper we presented a novel approach to address the problem of face pose estimation, by adopting simple and reliable Haar feature-based cascaded classifiers, together with the use of LWPR, an online incremental regression-based learning strategy. To experimentally evaluate the performance of our proposed approach, we acquired a large amount of real labelled images using a swarm of UAVs, and used these to perform emulation tests.

Acknowledgements

This research was funded by the Swiss National Science Foundation (SNSF) through the National Centre of Competence in Research (NCCR) Robotics. The authors participate in Project 4: *Distributed robotics*, Sub-project 4.5: *Symbiotic human-swarm cooperation*.

5. REFERENCES

[1] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2879–2886.

[2] Y. Pan, H. Zhu, and R. Ji, "3d head pose estimation for monocular image," in *Proc. of Intl. Conf, on Fuzzy Systems and Knowledge Discovery*, 2005, pp. 293–301.

[3] M. Demirkus, B. Oreshkin, J. J. Clark, and T. Arbel, "Spatial and probabilistic codebook template based head pose estimation from unconstrained environments," in *Proc. of IEEE Intl. Conf. on Image Processing (ICIP)*, 2011, pp. 573–576.

[4] H. T. Ho and R. Chellappa, "Automatic head pose estimation using randomly projected dense sift descriptors," in *Proc. of IEEE Intl. Conf. on Image Processing (ICIP)*, 2012, pp. 153–156.

[5] H. Ji, R. Liu, F. Su, Z. Su, and Y. Tian, "Robust head pose estimation via convex regularized sparse regression," in *Proc. of IEEE Intl. Conf. on Image Processing (ICIP)*, 2011, pp. 3617–3620.

[6] O. Williams, A. Blake, and R. Cipolla, "Sparse and semi-supervised visual mapping with the s^{3}gp," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2006, pp. 230–237.

[7] A. Ranganathan and M.-H. Yang, "Online sparse matrix gaussian process regression and vision applications," in *Proc. of European Conference on Computer Vision (ECCV)*, 2008, pp. 468–482.

[8] A. Asthana, S. Zafeiriou, S. Cheng, and M. Pantic, "Robust discriminative response map fitting with constrained local models," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2013, pp. 3444–3451.

[9] R. Valenti, Z. Yucel, and T. Gevers, "Robustifying eye center localization by head pose cues," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2009, pp. 612–618.

[10] T. Vatahska, M. Bennewitz, and S. Behnke, "Feature-based head pose estimation from images," in *Proc. of IEEE-RAS Intl. Conf. on Humanoid Robots (HUMANOIDS)*, 2007, pp. 330–335.

[11] G. Fanelli, M. Dantone, J. Gall, A. Fossati, and L. V. Gool, "Random forests for real time 3d face analysis," *International Journal of Computer Vision*, vol. 101, no. 3, pp. 437–458, 2013.

[12] M. A. Haj, "On partial least squares in head pose estimation: How to simultaneously deal with misalignment," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2602–2609.

[13] G. Fanelli, J. Gall, and L. V. Gool, "Real time head pose estimation with random regression forests," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 617–624.

[14] Y. Li, S. Gong, and H. Liddell, "Support vector regression and classification based multi-view face detection and recognition," in *Proc. of Intl. Conf. on Automatic Face and Gesture Recognition*, 2000, pp. 300–305.

[15] J. Nagi, H. Ngo, A. Giusti, L. M. Gambardella, J. Schmidhuber, and G. A. Di Caro, "Incremental learning using partial feedback for gesture-based human-swarm interaction," in *Proc. of the 21st IEEE Intl. Symp. on Robots and Human Interactive Communication (RO-MAN)*, 2012, pp. 898–905.

[16] M. Bonani, V. Longchamp, S. Magnenat, P. Retornaz, D. Burnier, G. Roulet, F. Vaussard, H. Bleuler, and F. Mondada, "The marxbot, a miniature mobile robot opening new perspectives for the collective-robotic research," in *Proc. of the IEEE/RSJ Intl. Conf. on Intelligent Robots and Systems (IROS)*, 2010, pp. 4187–4193.

[17] P. Viola and M. Jones, "Robust real-time face detection," *International Journal of Computer Vision*, vol. 57, no. 2, pp. 137–154, 2004.

[18] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. of IEEE Intl. Conf. on Computer Vision and Pattern Recognition (CVPR)*, 2001, pp. 511–518.

[19] S. Vijayakumar, A. D'souza, T. Shibata, J. Conradt, and S. Schaal, "Statistical learning for humanoid robots," *Autonomous Robots*, vol. 12, no. 1, pp. 55–69, 2002.

[20] H. Glaude, F. Akrimi, M. Geist, and O. Pietquin, "A non-parametric approach to approximate dynamic programming," in *Proc. of Intl. Conf. on Machine Learning and Applications (ICMLA)*, 2011, pp. 317–322.

[21] S. Klanke, S. Vijayakumar, and S. Schaal, "A library for locally weighted projection regression," *Journal of Machine Learning Research (JMLR)*, vol. 9, pp. 623–626, 2008.

[22] S. Vijayakumar, A. D'souza, and S. Schaal, "Incremental online learning in high dimensions," *Neural Computation*, vol. 17, no. 12, pp. 2602–2634, 2005.

978-1-4799-5751-4/14/$31.00 ©2014 IEEE