Head-pose Estimation In-the-Wild Using a Random Forest

Roberto Valle[1] José Miguel Buenaposada[2]
Antonio Valdés[3] Luis Baumela[1]

[1] Univ. Politécnica Madrid
{rvalle, lbaumela}@fi.upm.es

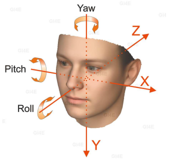
[2] Univ. Rey Juan Carlos
josemiguel.buenaposada@urjc.es

[3] Univ. Complutense Madrid
avaldes@ucm.es

July 14, 2016

Head-pose estimation

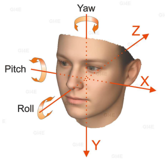
Predicting the relative orientation between the viewer and a target head



from the appearance in an image.

Head-pose estimation

Predicting the relative orientation between the viewer and a target head



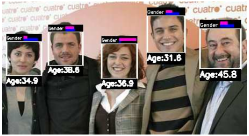
from the appearance in an image.

We consider the problem of estimating discretized yaw angles

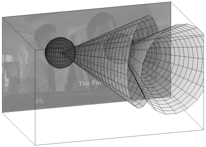
Applications

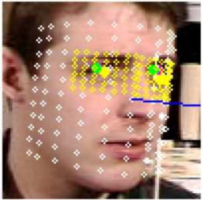
Preprocessing step for

- **Facial attributes**. Identity, age, gender, expression, . . .



- **HMI, FoA, Gaze.**



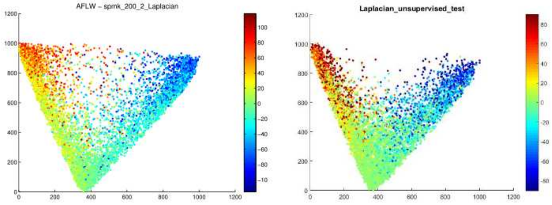


Depending on the estimation method

- **Subspace approaches**.
- **Flexible models.**
- **Classification.**
- **Regression.**

Depending on the estimation method

- **Subspace approaches**. Facial appearance changes lie on a low-dimesional manifold.
  [Sundararajan, AMFG15], [BenAbdelkader, ECCV10], [Balasubramanian, CVPR07]

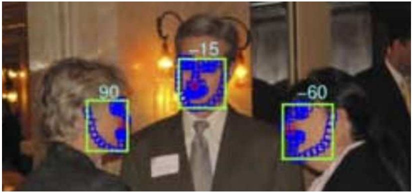


(b) AFLW (c) McGill Faces

- **Flexible models.**
- **Classification.**
- **Regression.**

Depending on the estimation method

- **Subspace approaches**. Facial appearance changes lie on a low-dimesional manifold.
- **Flexible models.** Fit a deformable model and estimate pose from the position of a set of landmarks.
  [Zhu, CVPR12]



- **Classification.**
- **Regression.**

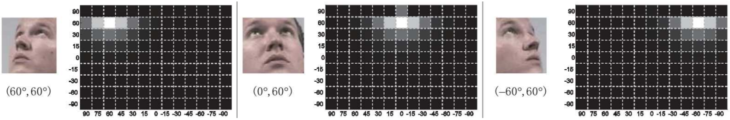
Depending on the estimation method

- **Subspace approaches**. Facial appearance changes lie on a low-dimesional manifold.
- **Flexible models.** Fit a deformable model and estimate pose from the position of a set of landmarks.
- **Classification.** Discretize poses in a group of classes. [Wu, PR08]

- **Regression.**

Depending on the estimation method

- **Subspace approaches**. Facial appearance changes lie on a low-dimesional manifold.
- **Flexible models.** Fit a deformable model and estimate pose from the position of a set of landmarks.
- **Classification.** Discretize poses in a group of classes.
- **Regression.** Estimate a continuous mapping function. [Haj, CVPR12], [Geng, CVPR14], [Hara, ECCV14]



Depending on the input data

- **Colour images**. [All previous]

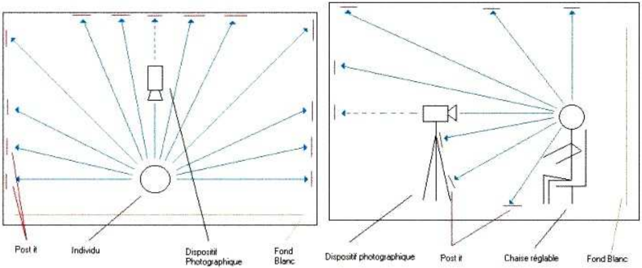
- **Depth images.** [Fanelli, IJCV 13]

Benchmarks

- **Faces "in the lab".**
  - **Pointing 04.**
  - **MultiPie.**
- **Faces "in the wild".**
  - **Annotated Facial Landmarks in the Wild (AFLW).**
  - **Annotated Faces in the Wild (AFW).**

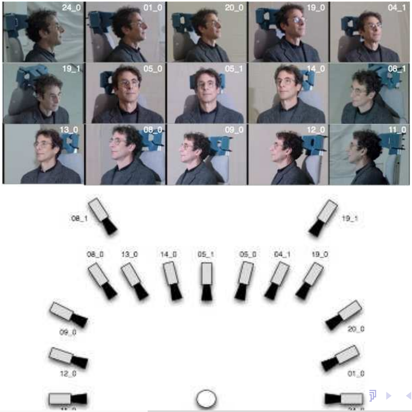
Benchmarks

- **Faces "in the lab".**
  - **Pointing 04.** 2790 images of 15 subjects spanning discrete yaw and pitch poses from -90° to 90° with 15° interval.



Benchmarks

- **Faces "in the lab".**
  - **Pointing 04.**
  - **MultiPie.** More than 750,000 images of 337 people, 15 view points, 19 illumination conditions, facial expressions.



Benchmarks

- **Faces "in the lab".**
  - **Pointing 04.**
  - **MultiPie.**
- **Faces "in the wild".**
  - **Annotated Facial Landmarks in the Wild (AFLW).** 25993 faces from *Flickr*, 59% female, 41% male. 380k manually annotated facial landmarks (21 point markup).

  - **Annotated Faces in the Wild (AFW).**

Benchmarks

- **Faces "in the lab".**
  - **Pointing 04.**
  - **MultiPie.**
- **Faces "in the wild".**
  - **Annotated Facial Landmarks in the Wild (AFLW).**
  - **Annotated Faces in the Wild (AFW).** 250 images with 468 challenging faces providing discrete yaw poses from -90° to 90° with 15° interval.

Best results

- **Faces “in the lab”.**

Method	Pointing-04 MAE
Stiefelhagen ICPRW04	9.5°
Haj CVPR12	6.56°
Hara ECCV14	5.29°
Geng CVPR14	4.24°

- **Faces “in the wild”.**

Best results

- **Faces "in the lab"**.
- **Faces "in the wild"**.

Method	AFLW MAE	AFW MAE
Sundararajan CVPR15	17.48°	17.20°

Head-pose classification

Our approach:

- Only estimate "yaw"
- Classification scheme.
- Discretize the range of yaw poses in steps of 15°.
- Estimate face orientation with a 13-class classifier.

Head-pose classification based on a Random Forest

- Ensemble of trees.
- Prediction determined by combining the outputs of all trees.
- For each tree, the leaf node provides a discrete distribution of head-pose.

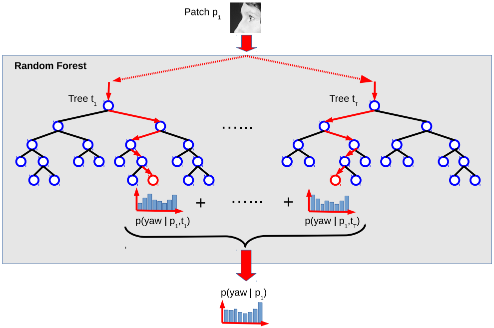


Image channels patches

- **Image channels.** We extract 38 channels $\mathtt{I}^{\alpha}$: gray-scale values, Sobel borders and 35 Gabor filters.

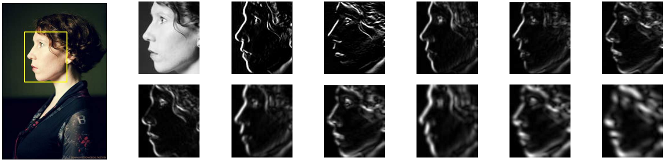
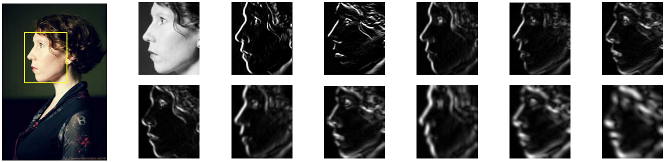
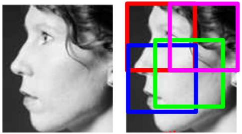


Image channels patches

- **Image channels.** We extract 38 channels $\mathtt{I}^{\alpha}$: gray-scale values, Sobel borders and 35 Gabor filters.



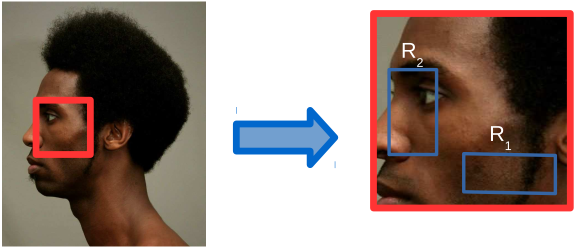
- **Image patches.** Randomly choose a set of square patches $\mathcal{P}_i = \{(\mathcal{I}_i, h_i)\}$, where $\mathcal{I}_i$ is the patch appearance and h_i is the pose.



Patch-based channel features

- Our features $\theta = (R_1, R_2, \alpha)$ are the difference between two rectangles, R_1 and R_2, within the patch in channel α.

$$f(p,\theta) = \frac{1}{|R_1|}\sum_{\mathbf{q}\in R_1} \mathtt{I}^{\alpha}(\mathbf{q}) - \frac{1}{|R_2|}\sum_{\mathbf{q}\in R_2} \mathtt{I}^{\alpha}(\mathbf{q})$$



Training Regression Forest

- Train each decision tree using a randomly selected set of patches from a random subset of the training faces.
- Optimize each weak learner by selecting the $\theta = (R_1, R_2, \alpha)$, from a random pool of candidates $\phi = (\theta, \tau)$, that maximizes the information gain

$$IG(\phi) = \mathcal{H}(\mathcal{P}) - \sum_{S \in \{L,R\}} \frac{|\mathcal{P}_S(\phi)|}{|\mathcal{P}|} \mathcal{H}(\mathcal{P}_S(\phi)), \quad (1)$$

where

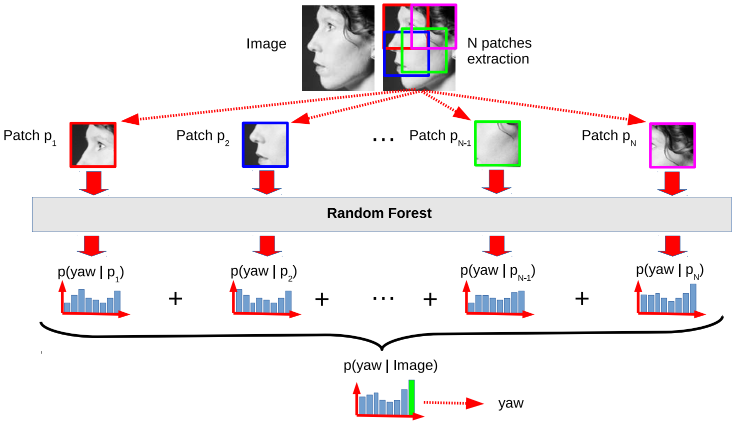
τ is the threshold over the feature value,

$\mathcal{P}_L(\phi) = \{\mathcal{P} | f(P, \theta) < \tau\}$,

$\mathcal{P}_R(\phi) = \mathcal{P} \setminus \mathcal{P}_L(\phi)$,

$\mathcal{H}(\mathcal{P}) = log(\sigma\sqrt{2\pi e})$.

Discrete pose estimation from face patches



Pose estimation algorithm

1. Detect face bounding box in $\mathtt{I}$.
2. Resize bounding box to $W \times H$ pixels, denoted $\mathtt{I}_r$.
3. Compute α channels from $\mathtt{I}_r$.
4. Extract from $\mathtt{I}_r$ the set of input patches $\mathcal{P}$ of size $N \times N$, with a stride of S pixels.
5. For each patch $p_i \in \mathcal{P}$:
    1. For each tree t_j in the forest:
        1. Input p_i to t_j.
        2. The leaf node of t_j reached by p_i provides a discrete distribution of the face orientation, $p(yaw|p_i, t_j)$.
    2. Compute the patch face pose distribution, $p(yaw|p_i) = \sum_j p(yaw|p_i, t_j)$.
6. Compute the final face pose distribution, $p(yaw|\mathtt{I}_r) = \sum_i p(yaw|p_i)$.

Algorithm configuration

- Resize face bounding box to 105×125 pixels.
- Forest with $T = 20$ trees each trained on a random set of images equally distributed by yaw angle.
- Extract 20 random patches of 61×61 pixels from face bounding box.
- Growing stops when depth reaches 15, or if there are less than 20 patches in a leaf.
- Select the best parameters from a pool of $\phi = 50000$ samples obtained from $\theta = 2000$ different combinations of $[\alpha$, R1, R2$]$ and $\tau = 25$ thresholds.
- The maximum random size of the subpatches defining the asymmetric areas R1 and R2 is set to be lower than a 75% of the patch size.
- Filter out leaves with a maximum variance threshold set to 400.

Databases

- Laboratory conditions evaluation
  - **Pointing-04**.
- Evaluation "in-the-wild".
  - **AFLW**.
  - **AFW**.

Qualitative results

Results for *Pointing-04* (top), *AFLW* (middle) and *AFW* (bottom) databases.

Green and blue lines indicate respectively estimated pose and ground truth.

Quantitative results in laboratory conditions

Our approach has a MAE close to the state-of-the-art.

Method	Pointing-04	
	MAE	Accuracy (0°)
Stiefelhagen ICPRW04	9.5°	52.0%
Haj CVPR12	6.56°	67.36%
Hara ECCV14	5.29°	-
Geng CVPR14	4.24°	73.30%
Our method	**7.84**°	**55.19%**

All three approaches with best results use holistic HOG-based face features.
In this constrained context, a global feature is more informative for estimating face pose than the set of local patches.

Quantivative results in real-world conditions

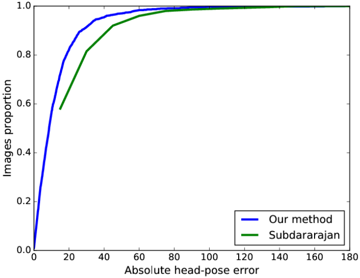
Our approach achieves the best performance.

Method	AFLW		AFW	
	MAE	Acc ($\leq 15^\circ$)	MAE	Acc ($\leq 15^\circ$)
Haj CVPR12	-	-	-	78.7%
Zhu CVPR12	-	-	-	81.0%
Sundararajan CVPR15	17.48°	58.05%	17.20°	58.33%
Our method	**12.26°**	**72.57%**	**12.50°**	**83.54%**

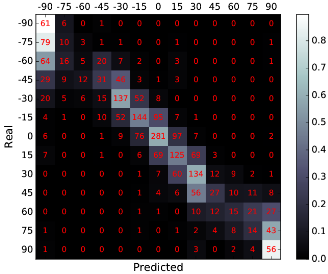
Our approach can deal with challenging in-the-wild conditions, such as the presence of occlusions, illumination changes or facial expressions.

Results AFLW

Cumulative error

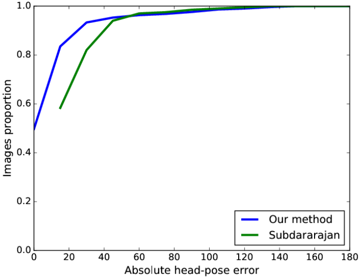


Confusion matrix

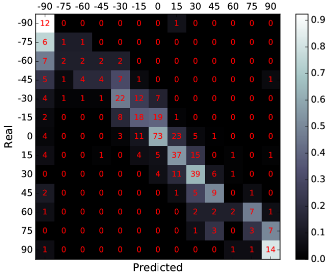


Results AFW

Cumulative error



Confusion matrix



Conclusions

We have presented an algorithm to estimate head-pose yaw angle in unconstrained settings

- Performs behind the state-of-the-art in laboratory conditions and better using “in the wild” databases.
- Local features provide good results in realistic imaging conditions.
- Achieves 80 FPS (12ms per image). It outperforms its competitors in terms of computational requirements.

Future use for estimation of facial attributes.