Robust Facial Landmark Detection via Recurrent Attentive-Refinement Networks

Shengtao XIAO, Jiashi FENG, Junliang XING, Hanjiang LAI, Shuicheng YAN, Ashraf KASSIM
Problem Introduction

• Obtain face shape by locating pre-defined facial landmarks.

• Challenges: face occlusions, pose variations, expressions, etc.

• Solutions: cascaded face shape regression

\[
\varphi(I, S^0) \rightarrow R^1 \xrightarrow{\Delta S^1} S^1 \xrightarrow{\cdots} \varphi(I, S^{t-1}) \rightarrow R^t \xrightarrow{\Delta S^t} S^t
\]
Recurrent Attentive-Refinement (RAR) Network

- Deep Feature Learning
- Robust Shape Initialization
- Recurrent-Attentive Refinement
  - Attention module
  - Refinement module
Background

- CNN model with Shape-Indexed Pooling (SIP)

- RNN model and Long Short-Term Memory (LSTM)
RAR Networks

A) Deep Feature Learning

Direct Prediction $S_d$
RAR Networks

A). Deep feature extraction, landmark regression and robust initialization.
RAR Networks

A). Deep feature extraction, landmark regression and robust initialization.
B). RAR sequentially refines the landmark estimation.
RAR Networks

A). Deep feature extraction, landmark regression and robust initialization.
B). RAR sequentially refines the landmark estimation.
C). An attention model in RAR for adaptively selecting key landmark points.
RAR Networks: Deep Feature Learning

- Modified VGG19 Network + two Deconvolution layers to ensure pixel-to-pixel correspondence
RAR Networks: Deep Feature Learning

- Modified VGG19 Network + two Deconvolution layers to ensure pixel-to-pixel correspondence

- SoftMax regression loss on conv8
RAR Networks: Deep Feature Learning

- Modified VGG19 Network + two Deconvolution lby selecting location of maximum response from \(v\)-th channel of \(\text{conv8}\)

- SoftMax regression loss on \(\text{conv8}\)

- Directly estimate landmark location \(S_d v d v d v d v d v d v\) by selecting location of maximum response from \(v\)-th channel of \(\text{conv8}\)
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion

Robust Initial Shape Selection:

\[ S_0 = \arg \min_S ||S - S_d||, \text{ s.t. } S \in \mathcal{F} \]
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion

Robust Initial Shape Selection:

$$S_0 = \arg\min_S ||S - S_d||, \text{ s.t. } S \in \mathcal{F}$$

$$S_0 = \arg\min_{S,c} ||S - S_d||_0 + \lambda ||c||_0, \text{ s.t. } S = \sum_i^{m} c_i S_i$$

GT Shapes
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion

Robust Initial Shape Selection:

\[ S_0 = \arg \min_S ||S - S_d||, \text{ s.t. } S \in \mathcal{F} \]

\[ S_0 = \arg \min_{S,c} ||S - S_d||_0 + \lambda ||c||_0, \text{ s.t. } S = \sum_i c_i S_i \]

Solve: get K representative shapes via K-meanings clustering + RANSAC method to filter out significant outliers
RAR Networks: Robust Shape Initialization

However, detected shape is sensitive to occlusion

Robust Initial Shape Selection:

\[
S_0 = \arg \min_S ||S - S_d||, \text{ s.t. } S \in \mathcal{F}
\]

\[
S_0 = \arg \min_{S,c} ||S - S_d||_0 + \lambda ||c||_0, \text{ s.t. } S = \sum_{i} c_i S_i
\]

Solve: get K representative shapes via K-meanings clustering + RANSAC method to filter out significant outliers
RAR Networks: Attention Module
RAR Networks: Attention Module

• A-LSTM (attention module) selects attention center with top confidence at each recurrent stage

\[ C^* = \arg\max_{c \in \{1, \ldots, L\}} \text{A-LSTM}(\Phi_a(I_t, \hat{S}_t); W_a, c) \]
RAR Networks: Attention Module

• $R_a a a R_a$

• A-LSTM (attention module) selects attention center with top confidence at each recurrent stage:

$$C^* = \arg\max_{c \in \{1, \ldots, L\}} A\text{-LSTM}(\Phi_a(I_t, S_t), W_a, c)$$

• A typical attention center is selected based on maximize reward $R_a$

$$R_a = \sum_{t=1}^{\infty} \eta^{t-1} R(\hat{S}_{t-1}, \hat{S}_t)$$

Update: $A_{t+1} S_t$
RAR Networks: Attention Module

- $\mathcal{R} a a a \mathcal{R} a$
- A-LSTM (attention module) selects attention center with top confidence at each recurrent stage:
  
  $$c^* = \arg \max_{c \in \{1, \ldots, L\}} \mathcal{A-LSTM}(\mathcal{A}(l_t, S_t), W_a, c)$$

- A typical attention center is selected based on maximize reward $\mathcal{R} a$
  
  $$\mathcal{R}_a = \sum_{t=1}^{\infty} \eta^{t-1} R(\hat{S}_{t-1}, \hat{S}_t)$$

  $$R(\hat{S}_{t-1}, \hat{S}_t) = ||\Gamma_t \Delta S_{t-1}||^2_2 - ||\Gamma_t \Delta S_t||^2_2$$

  $$\text{Update} = \Delta_k S_t$$
RAR Networks: Attention Module

- $\mathcal{R}_a \mathcal{a} \mathcal{a} \mathcal{R}_a$

- A-LSTM (attention module) selects attention center with top confidence at each recurrent stage.

$$\mathcal{C}^* = \arg \max_{c \in \{1, \ldots, L\}} \mathcal{A}-\text{LSTM}(\Phi_a(t, S_t), W_a, c)$$

- A typical attention center is selected based on maximize reward $\mathcal{R}_a$

$$\mathcal{R}_a = \sum_{t=1}^{\infty} \eta^{t-1} R(\hat{S}_{t-1}, \hat{S}_t)$$

$$R(\hat{S}_{t-1}, \hat{S}_t) = || \Gamma_t \Delta S_{t-1} ||_2^2 - || \Gamma_t \Delta S_t ||_2^2$$

$$\Gamma_t = [\gamma_t^1, \gamma_t^2, \ldots, \gamma_t^L], \text{ with } \gamma_t^l = \kappa \exp \left( -\frac{|| \hat{S}_t^l - \hat{S}_t^c^* ||_2^2}{4D_t^2} \right)$$
RAR Networks: Refinement Module

- Feature re-weighting based on distance to attention center:

\[ \Phi_r(I_t, \hat{S}_t) = [\gamma_t^1 \phi_t^1, \gamma_t^2 \phi_t^2, ..., \gamma_t^L \phi_t^L] \]
RAR Networks: Refinement Module

- Feature re-weighting based on distance to attention center:
  \[ \Phi_r(I_t, \hat{S}_t) = [\gamma_1^t \phi_1^t, \gamma_2^t \phi_2^t, ..., \gamma_L^t \phi_L^t] \]

- Refinement Module to get shape update such:
  \[ \mathcal{L}_R^t = || \Gamma_t(\Delta_R S_t) - \Delta S_t ||_2^2 \]
  \[ \Delta R S_t \] is the R-LSTM output \[ \Delta R S_t = \alpha \Gamma_t R-LSTM(\Phi_r) \]
RAR Networks: Refinement Module

- Feature re-weighting based on distance to attention center:
  \[ \Phi_r(I_t, \hat{S}_t) = [\gamma_t^1 \phi_t^1, \gamma_t^2 \phi_t^2, ..., \gamma_t^L \phi_t^L] \]

- Refinement Module to get shape update such:
  \[ \mathcal{L}_R^t = || \Gamma_t (\Delta_R S_t - \Delta S_t) ||_2^2 \]

\( \Delta_R S_t \) is the R-LSTM output \( \Delta_R S_t = \alpha \Gamma_t R-LSTM(\Phi_r) \)

Overall Training Objective of RAR:

\[
\sum_{t=1}^{T} \sum_{n=1}^{N} -\eta^{t-1} R_a(\hat{S}_{t-1,n}, \hat{S}_{t,n}) + \mathcal{L}_{R,n}^t
\]

- Attention Loss
- Regression Loss
## Results on 300W

<table>
<thead>
<tr>
<th>Methods</th>
<th>300-W Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Common</td>
</tr>
<tr>
<td>Zhu et.al [2012]</td>
<td>8.22</td>
</tr>
<tr>
<td>RCPR [Burgos,2013]</td>
<td>6.18</td>
</tr>
<tr>
<td>SDM [Xiong,2013]</td>
<td>5.57</td>
</tr>
<tr>
<td>LBF [Ren,2014]</td>
<td>4.95</td>
</tr>
<tr>
<td>LBF Fast [Ren,2014]</td>
<td>5.38</td>
</tr>
<tr>
<td>CFAN[Zhang, 2014]</td>
<td>5.50</td>
</tr>
<tr>
<td>CFSS [Zhu, 2015]</td>
<td>4.73</td>
</tr>
<tr>
<td>Ours (RAR)</td>
<td>4.12</td>
</tr>
</tbody>
</table>

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015
### Results on 300W

<table>
<thead>
<tr>
<th>Methods</th>
<th>300-W Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Common</td>
</tr>
<tr>
<td>Zhu et. al [2012]</td>
<td>8.22</td>
</tr>
<tr>
<td>RCPR [Burgos,2013]</td>
<td>6.18</td>
</tr>
<tr>
<td>SDM [Xiong,2013]</td>
<td>5.57</td>
</tr>
<tr>
<td>LBF [Ren,2014]</td>
<td>4.95</td>
</tr>
<tr>
<td>LBF Fast [Ren,2014]</td>
<td>5.38</td>
</tr>
<tr>
<td>CFAN [Zhang, 2014]</td>
<td>5.50</td>
</tr>
<tr>
<td>CFSS [Zhu, 2015]</td>
<td>4.73</td>
</tr>
<tr>
<td>Ours (RAR)</td>
<td>4.12</td>
</tr>
</tbody>
</table>

---

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012  
RCPR: Robust face landmark estimation under occlusion. ICCV 2013  
SDM: Supervised descent method and its applications to face alignment. CVPR 2013  
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014  
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014  
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015
Results on 300W

<table>
<thead>
<tr>
<th>Methods</th>
<th>300-W Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Common</td>
</tr>
<tr>
<td>Zhu et.al [2012]</td>
<td>8.22</td>
</tr>
<tr>
<td>RCPR [Burgos,2013]</td>
<td>6.18</td>
</tr>
<tr>
<td>SDM [Xiong,2013]</td>
<td>5.57</td>
</tr>
<tr>
<td>LBF [Ren,2014]</td>
<td>4.95</td>
</tr>
<tr>
<td>LBF Fast [Ren,2014]</td>
<td>5.38</td>
</tr>
<tr>
<td>CFAN[Zhang, 2014]</td>
<td>5.50</td>
</tr>
<tr>
<td>CFSS [Zhu, 2015]</td>
<td>4.73</td>
</tr>
<tr>
<td>Ours (RAR)</td>
<td>4.12</td>
</tr>
</tbody>
</table>

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015
Results on 300W

<table>
<thead>
<tr>
<th>Methods</th>
<th>300-W Dataset</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Common</td>
</tr>
<tr>
<td>Zhu et.al [2012]</td>
<td>8.22</td>
</tr>
<tr>
<td>RCPR [Burgos,2013]</td>
<td>6.18</td>
</tr>
<tr>
<td>SDM [Xiong,2013]</td>
<td>5.57</td>
</tr>
<tr>
<td>LBF [Ren,2014]</td>
<td>4.95</td>
</tr>
<tr>
<td>LBF Fast [Ren,2014]</td>
<td>5.38</td>
</tr>
<tr>
<td>CFAN[Zhang, 2014]</td>
<td>5.50</td>
</tr>
<tr>
<td>CFSS [Zhu, 2015]</td>
<td>4.73</td>
</tr>
<tr>
<td>Ours (RAR)</td>
<td>4.12</td>
</tr>
</tbody>
</table>

Zhu: Face detection, pose estimation, and landmark localization in the wild. CVPR 2012
RCPR: Robust face landmark estimation under occlusion. ICCV 2013
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
LBF: Face alignment at 3000 fps via regressing local binary features. ECCV 2014
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015
Results on COFW and AFLW

<table>
<thead>
<tr>
<th>Methods</th>
<th>Normalized ME</th>
<th>Failure Rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCPR</td>
<td>8.50</td>
<td>20.00%</td>
</tr>
<tr>
<td>HPM</td>
<td>7.46</td>
<td>13.24%</td>
</tr>
<tr>
<td>RPP</td>
<td>7.52</td>
<td>16.20%</td>
</tr>
<tr>
<td>TCDCN</td>
<td>8.05</td>
<td>-</td>
</tr>
<tr>
<td>RAR</td>
<td>6.03</td>
<td>4.14%</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Methods</th>
<th>Normalized ME</th>
</tr>
</thead>
<tbody>
<tr>
<td>RCPR</td>
<td>11.6</td>
</tr>
<tr>
<td>SDM</td>
<td>8.50</td>
</tr>
<tr>
<td>CFAN</td>
<td>10.95</td>
</tr>
<tr>
<td>TCDCN</td>
<td>7.60</td>
</tr>
<tr>
<td>RAR</td>
<td>7.23</td>
</tr>
</tbody>
</table>

RCPR: Robust face landmark estimation under occlusion. ICCV 2013
RPP: Regional Predictive Power. TIP 2015.
TCDCN: Task constraint deep convolutional nets. PAMI 2015.
SDM: Supervised descent method and its applications to face alignment. CVPR 2013
CFAN: Coarse-to-ne auto-encoder networks(cfan) for real-time face alignment. ECCV2014
CFSS: Face alignment by coarse-to-fine shape searching. CVPR 2015
## Comparison Studies

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conv8</th>
<th>Mean Shape</th>
<th>Random Shape</th>
<th>Direct</th>
<th>Robust Initialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-W</td>
<td>6.24</td>
<td>5.26</td>
<td>5.22</td>
<td>6.66</td>
<td>4.94</td>
</tr>
<tr>
<td>COFW</td>
<td>30.14</td>
<td>6.24</td>
<td>6.12</td>
<td>11.52</td>
<td>6.03</td>
</tr>
<tr>
<td>AFLW</td>
<td>8.14</td>
<td>7.36</td>
<td>7.42</td>
<td>8.15</td>
<td>7.23</td>
</tr>
</tbody>
</table>

**Mean Shape**: Prediction from conv8  
**Direct**: RAR trained with initial shape as conv8  
**Mean Shape**: RAR trained with mean shape as initial shape  
**Random Shape**: RAR trained with random shape as initial shape  
**Robust**: RAR trained with the proposed robust initialization
Comparison Studies

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conv8</th>
<th>Mean Shape</th>
<th>Random Shape</th>
<th>Direct</th>
<th>Robust Initialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-W</td>
<td>6.24</td>
<td>5.26</td>
<td>5.22</td>
<td>6.66</td>
<td>4.94</td>
</tr>
<tr>
<td>COFW</td>
<td>30.14</td>
<td>6.24</td>
<td>6.12</td>
<td>11.52</td>
<td>6.03</td>
</tr>
<tr>
<td>AFLW</td>
<td>8.14</td>
<td>7.36</td>
<td>7.42</td>
<td>8.15</td>
<td>7.23</td>
</tr>
</tbody>
</table>

Conv8: Prediction from conv8
Direct: RAR trained with initial shape as conv8
Mean Shape: RAR trained with mean shape as initial shape
Random Shape: RAR trained with random shape as initial shape
Robust: RAR trained with the proposed robust initialization
## Comparison Studies

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Conv8</th>
<th>Mean Shape</th>
<th>Random Shape</th>
<th>Direct</th>
<th>Robust Initialization</th>
</tr>
</thead>
<tbody>
<tr>
<td>300-W</td>
<td>6.24</td>
<td>5.26</td>
<td>5.22</td>
<td>6.66</td>
<td>4.94</td>
</tr>
<tr>
<td>COFW</td>
<td>30.14</td>
<td>6.24</td>
<td>6.12</td>
<td>11.52</td>
<td>6.03</td>
</tr>
<tr>
<td>AFLW</td>
<td>8.14</td>
<td>7.36</td>
<td>7.42</td>
<td>8.15</td>
<td>7.23</td>
</tr>
</tbody>
</table>

**Mean Shape**: Prediction from conv8

**Random Shape**: RAR trained with initial shape as conv8

**Direct**: RAR trained with mean shape as initial shape

**Attentionless**

**Conv8**: RAR trained with random shape as initial shape

**Robust**: RAR trained with the proposed robust initialization
Attention Center Selection Frequencies
Attention Center Selection Frequencies

Stage 1 to Stage 5

Stage 6 to Stage 10
Attention Center Selection Frequencies

Stage 1 to Stage 5

Stage 6 to Stage 10

Stage 11 to Stage 15
Sample Attentive Refinement

Iter 01

Iter 01
Sample Attentive Refinement
Sample Attentive Refinement
Sample Attentive Refinement
Sample Attentive Refinement
Sample Attentive Refinement
Sample Attentive Refinement
Sample Attentive Refinement
Q&A

Poster Session: O-1A-04