Self-supervised learning of a facial attribute embedding from video

Summary

1. Learns an embedding that encodes information about head pose, facial landmarks and facial expression without labelled supervision.

2. Given a src frame & tgt frame, the pipeline must learn to reconstruct the tgt frame by learning a bilinear sampler on the src frame.

3. The method outperforms other SSL method given the same training data.

4. Reason given: "as parts of the face move tgt (e.g. an eyebrow raise), the emb. must learn to encode info. about facial features and thereby encode expression".

Single-src frame architecture

1. Given src frame s and target frame t
2. $v_t = f(t)$ and $v_s = f(s)$ where f is an encoder
3. v_t and v_s are concatenated & fed to a decoder
4. The decoder learns a mapping g from a concatenated embedding to a bilinear grid sampler.
5. The bilinear grid sampler samples from the scr frame to create a new frame $s' = g(v_t, v_s)(s)$.
6. Precisely, g predicts $(\delta x, \delta y)$ and $s'(x, y) = s(x + \delta x, y + \delta y)$
7. The network is trained to minimize $\alpha(s', t) = \| t - s' \|_1$.

Multi-source frame architecture

1. Additional frame can be used to improve the quality of the learned embedding.
2. for each scr frame, an additional decoder predicts a confidence heatmap C_i.
3. The confidence heatmap C_i are combined pixel-wise for each scr frame s_i using a softmax operation.

Curriculum Strategy

1. Using progressively more difficult samples is important for successful learning.
2. The loss computed for samples in a batch is used to ranked them.
3. Initially, the loss is backproped only on samples with smaller loss.
4. When the loss on the validation set plateaus, training is performed with more challenging samples.