Automatic Face Aging in Videos via Deep Reinforcement Learning (Summary)

Authors :
Chi Nhan Duong [1], Khoa Luu [2], Kha Gia Quach [1], Nghia Nguyen [2], Eric Patterson [3], Tien D. Bui [1], Ngan Le [4]

1 Computer Science and Software Engineering, Concordia University, Canada
2 Computer Science and Computer Engineering, University of Arkansas, USA
3 School of Computing, Clemson University, USA
4 Electrical and Computer Engineering, Carnegie Mellon University, USA

1fdcnhan, kquachg@ieee.org, bui@encs.concordia.ca, 2fkhoaluu, nhnguyeng@uark.edu,
3ekp@clemson.edu, 4thihoanl@andrew.cmu.edu

Introduction:

Age-related technologies generally address areas of age prediction and age progression. For the age recognition technologies, the model learns some aging-related features which are observed at several stages of age over the face, based on these features the model tries to predict when given a facial photograph whereas the age progression algorithms are more complex and learn more complex features that can describe the process of aging.

This paper discusses a method that models facial structures and the longitudinal face-aging process of given subjects coherently across video frames. They have optimized their method using long term reward, Reinforcement learning using deep feature extraction from Deep Convolutional Neural Network. The approach modeled the age transformation as optima selection using CNN. Instead of applying the image-based age progression separately the proposed method exploits the temporal relationship between two consecutive frames and because of this property, the model facilitates in maintaining consistency of aging information embedded into each frame. In the proposed structure a smoother synthesis can be produced along with the visual fidelity of aging data.

Previous Work:

In the field of age progression, there have been mainly two key research directions for both conventional computer vision and deep learning methods, which are one-shot synthesis and multiple-shot synthesis. These techniques attempt to discover aging patterns demonstrated over individuals or populations. In one shot synthesis, a new face at the target aegis directly synthesized via inferring the relationships between training images and their corresponding age labels and then applying them to generate the aged likeness. In multiple-shot synthesis, the longitudinal aging process is decomposed into multiple steps of aging effects. These methods are built on the facial aging transformation between two consecutive executive groups. Then finally the progressed faces from one age group to another age group are synthesized step by step until they reach the target age. The previous approaches can be divided into model-based, a reconstruction based, prototyping based, and deep learning-based. Model-based approaches aim at modeling both the shape and texture of facial images using the parameterization method, then learning to change these parameters via an aging function. Reconstruction-based approaches proposed to model personalized aging patterns by attempting to preserve distinct facial features of an individual through the aging process. In Deep-Learning based approaches, temporal and Spatial Restricted Boltzmann Machines (TRBM) were introduced in, to represent the non-linear aging process, with geometry constraints, and to model a sequence of reference

faces as well as wrinkles of adult faces. A Recurrent Neural Network (RNN) with a two-layer Gated Recurrent Unit (GRU) was employed to approximate aging sequences and finally in prototype-based approaches.

Dataset:

The authors collected a video dataset for temporal aging evaluations with 100 videos of celebrities. Each video clip consists of 200 frames. Age annotations were estimated using the year of the interview session versus the year of birth of the individual.

Approach:

1. The aging algorithm is formulated as the sequential decision-making process from a goal-oriented agent while interacting with the temporal visual environment.
2. The agent integrates related information of the current and previous frames then modifies action accordingly.
3. The agent receives a scalar reward at each time-step to maximize the total long-term aggregate of rewards, emphasizing the effective utilization of temporal observations in computing the aging transformation employed on the current frame.
4. The goal is to learn the synthesis function which maps the image pair at time step t to zero time step.
5. **Feature Embedding:** Main aim for this is leaning and embedding function to map the image pair at time step t into its latent representation which indirectly produces high quality synthesized images. The chosen function should produce feature representation with two main properties i.e. linearly separable and detail preserving.
6. **Manifold Traversing:** To interpret age progression as linear traversal for younger age region.
    - To compute the function containing only aging effects without the presence of other factors like identity, pose, etc., they exploit the relationship in terms of the aging changes between the nearest neighbors of image vector into two age groups.

    $$\Delta^{\mathbf{x}^t|\mathbf{X}^{1:t-1}} = \frac{1}{K}\left[\sum_{\mathbf{x}\in\mathcal{N}_o^t} \mathcal{F}_1(\mathcal{A}(\mathbf{x},\mathbf{x}_y^t)) - \sum_{\mathbf{x}\in\mathcal{N}_y^t} \mathcal{F}_1(\mathcal{A}(\mathbf{x},\mathbf{x}_y^t))\right]$$

    Where Delta is the temporal difference between the images frames at time-set t and 0. F is the feature embedding and $A(x; x_t^y)$ denotes a face-alignment operator, K is constant.
    - **Deep RL for Neighbour Selection:** To identify the nearest image frame they have used criteria names closeness criteria like such as distance in feature domain, or several matched attributes.
    - The drawback of this method is they are unable to maintain visually cohesive age information across video frames as the mentioned criteria arent frame independent.
    - Thus RL framework is proposed as a sequential decision-making process to maximize the temporal reward estimated by the consistency between the neighbor sets.
    - The agent policy framework will iteratively analyze the role of each neighbor of both young and old age combinations with relationships to determine new neighbors.
    - A new neighbor is considered to be appropriate when it is sufficiently similar to the older one, i.e. it maintains the age consistency between 2 frames.
    - Whenever a new neighbor is selected, the neighbor sets gets updated and the model receives the reward-based on the estimation of similarity of embedded aging information between 2 frames.
    - Thus, the agent can iteratively explore an optimal route for selecting neighbors to maximize the long-term reward.
    - The policy, actions, reward function, equation to synthesis from features, etc. are given as follows
    - **Actions:** Defined as selecting the new neighbor for the current frame such that with this new sample added to the neighbor sets of the current frame
    - **Policy Network:**

$$\mathbf{u}_i^t = \left[\delta_{\mathcal{F}_1}^{\text{pool5}}(\mathbf{x}_y^t, \mathbf{x}_y^{t-1}), \mathcal{F}_1^{\text{pool5}}(\mathbf{z}_i^{t-1})\right]$$
$$\mathbf{v}_i^t = \left[d\left((\mathcal{N}^t)_i, \mathbf{x}_y^t\right), d\left(\bar{\mathcal{N}}^t, \mathbf{x}_y^t\right), \mathbf{M}_i\right]$$

i: time step
Delta: the temporal difference between the images frames at time-set t and 0
F: Embedding Function
N: Samples
x: frame
d: distance between two frames

- **Reward Function:**

$$r_i^t = \frac{1}{\| \Delta_{i,\mathcal{A}(\cdot,\mathbf{x}_y^t)}^{\mathbf{x}^t|\mathbf{X}^{1:t-1}} - \Delta_{\mathcal{A}(\cdot,\mathbf{x}_y^t)}^{\mathbf{x}^{t-1}|\mathbf{X}^{1:t-2}} \| + \epsilon}$$

r: reward

- **Model Learning:** The training objective is to maximize the sum of the reward signals. The recurrent policy network with the REINFORCE algorithm guided by the reward given at each time step.
- **Synthesizing from the features:**

$$\mathbf{x}_o^{t*} = \arg\min_{\mathbf{x}} \frac{1}{2} \| \mathcal{F}_1(\mathbf{x}_o^t) - \mathcal{F}_1(\mathbf{x}) \|_2^2 + \lambda_{V^\beta} R_{V^\beta}(\mathbf{x})$$

After the neighbor sets are selected, the feature embeddings are given as the above function. R_V represents the Total Variation regularizer encouraging smooth transitions between pixel values.

Results:

1. **Database**: The proposed approach is trained and evaluated using training and testing databases that are not overlapped.
2. The neighbor sets are constructed using a largescale database composing face images AGFW-v2 and LFW-GOOGLE are used.
3. Then Policy network is trained using videos from 300-VW.
4. **Implementation:**
   Images from AGFWv2 and LFW-GOOGLE are combined and divided into 11 age groups from 10 to 65 with the age span of five years.
5. **Model Structure and Training:**
   - Employed a neural network with two hidden layers of 4096 and 2048 hidden units
   - Activation Function: Relu(for each hidden layer)
   - The videos from 300-VW are used to train the policy network
6. **Computation Time:**
   Then Policy network is trained using videos from 300-VW.
   System: Intel i7-6700, CPU@3.4GHz with an NVIDIA GeForce TITAN X GPU.
7. **Age Progression:** It demonstrates the validity of the approach for robustly and consistently synthesizing age-progressed faces across consecutive frames of input videos.
   - Age Progression in frontal and off-angle faces: the major changes between frames come from facial expressions and movements of the mouth and lips), or off-angle faces (i.e. more challenging
   - due to the pose effects in the combination of other variations)

- Wrinkles of soft-tissue areas (i.e. under the subject's eyes; around the cheeks and mouth) are coherent robust between consecutive synthesized frames.
- Aging consistency: They adopted the average inverted reward of all frames for each synthesis video. They firstly compute the optical flow, i.e. an estimation of image displacements, between frames of each video to estimate changes in pixels through time. Then we evaluate the differences (L2-norm) between the flows of the original versus synthesis videos.
- policy network has consistently and robustly shown its role in maintaining an appropriate aging amount embedded to each frame, and, therefore, producing smoother synthesis across frames in the output videos

Conclusion:

They have optimized their method using long term reward, Reinforcement learning using deep feature extraction from Deep Convolutional Neural Network. The model inherits the strengths of both recent advances of deep networks and reinforcement learning techniques to synthesize aged faces of given subjects both plausibly and coherently across video frames. The approach modeled the age transformation as optima selection using CNN. Instead of applying the image-based age progression separately the proposed method exploits the temporal relationship between two consecutive frames and because of this property, the model facilitates in maintaining consistency of aging information embedded into each frame. In the proposed structure a smoother synthesis can be produced along with the visual fidelity of aging data. Their method can generate age-progressed facial likenesses in videos with consistently aging features across frames. Moreover, our method guarantees the preservation of the subject's visual identity after synthesized aging effects.