中国科学: 信息科学 2019 年 第 49 卷 第 4 期: 450–463
SCIENTIA SINICA Informationis
论文

基于编解码网络的多姿态人脸图像正面化方法

徐海月[1,2], 姚乃明[1,2], 彭晓兰[1,2], 陈辉[1,2*], 王宏安[1,2,3]

1. 中国科学院软件研究所人机交互北京市重点实验室, 北京 100190
2. 中国科学院大学, 北京 100049
3. 中国科学院软件研究所计算机科学国家重点实验室, 北京 100190
* 通信作者. E-mail: chenhui@iscas.ac.cn

收稿日期: 2018–05–26; 接受日期: 2018–06–25; 网络出版日期: 2019–04–11

国家重点研发计划项目 (批准号: 2016YFB1001405)、国家自然科学基金项目 (批准号: 61661146002) 和中国科学院前沿科学重点研究计划项目 (批准号: QYZDY-SSW-JSC041) 资助

摘要 多姿态人脸图像正面化可以缓解头部姿态变化对人脸分析任务的影响. 以往直接从多姿态人脸图像合成正面人脸图像的方法存在细节特征缺失的问题. 针对这一问题, 本文提出一种基于编解码网络的多姿态人脸图像正面化方法 —— 多任务卷积编解码网络 (MCEDN). 该方法引入正面基础特征网络合成正面人脸基础特征, 并在此基础上融合编码网络提取的多姿态人脸局部特征进行细节补偿, 最终合成更加清晰的正面人脸图像. 利用多任务学习机制建立端到端模型, 统一局部特征提取、正面基础特征解析、正面图像合成 3 个模块, 通过共享参数提升整个模型的效果. 与已有方法对比, MCEDN 在多个数据集上都可以合成结构稳定、细节清晰的正面人脸图像. 我们直接使用合成的正面人脸图像进行人脸识别和表情识别, 识别准确率达到先进水平, 这表明 MCEDN 可以有效保留人脸细节特征, 支持人脸分析任务.

关键词 人脸正面化, 卷积神经网络, 编解码网络, 多任务学习, 人脸识别, 表情识别

1 引言

随着人工智能技术的发展, 多姿态人脸分析技术逐渐成为研究热点. 然而人脸图像中的头部姿态变化会造成面部细节特征的缺失, 削弱模型的鲁棒性. 多姿态人脸图像正面化正是缓解这一问题的方法之一. 多姿态人脸图像正面化通常采用特定算法通过多姿态人脸图像获取正面姿态的人脸图像 [1,2]. 这类方法的合成结果可以为不同的图像分析任务提供帮助, 但是现有的方法只针对近正面姿态变化具有较好的修正效果, 对于较大角度姿态, 由于人脸表观缺损过大, 修正后的正面图像效果不理想. 因此, 研究保留更多细节的多姿态人脸图像正面化方法是十分必要的.

引用格式: 徐海月, 姚乃明, 彭晓兰, 等. 基于编解码网络的多姿态人脸图像正面化方法. 中国科学: 信息科学, 2019, 49: 450–463, doi: 10.1360/N112018-00060
Xu H Y, Yao N M, Peng X L, et al. A multi-pose face frontalization method based on encoder-decoder network (in Chinese). Sci Sin Inform, 2019, 49: 450–463, doi: 10.1360/N112018-00060

针对多姿态人脸图像正面化如何保留更多细节的问题, 本文提出一种基于深度卷积编解码网络的多姿态人脸图像正面化方法. 该方法主要包含以下特点: (1) 引入正面基础特征网络合成正面人脸基础特征; (2) 融合多姿态人脸局部特征弥补正面人脸基础特征缺失的面部细节, 合成更加清晰的正面人脸图像; (3) 采用多任务学习机制 [3] 建立端到端模型, 统一局部特征提取、正面基础特征合成、正面图像合成 3 部分, 通过共享参数提升模型性能.

实验结果表明, 本文采用的 MCEDN 模型在表情识别、人脸识别任务上的识别准确率均为最高, 且姿态的角度越大, 识别准确率的提升程度越高, 这表明 MCEDN 模型可以从其他姿态人脸图像合成正面图像, 且合成的图像保留了丰富的人脸细节信息, 可以用于识别类任务.

2 相关工作

人脸图像中头部姿态变化引起的面部表观损失对于人脸识别、表情识别等依赖人脸表观特征的人脸分析任务影响重大, 为了从多姿态人脸图像获得正面人脸图像, 众多研究工作被提出, 这些工作大致可以分为基于三维模型的技术和基于多姿态二维图像的技术.

多姿态人脸图像与正面人脸图像的区别可以视为不同投影造成的形变扭曲和自遮挡. 因而, 可以通过恢复形变合成正面人脸图像. Zhu 等 [4] 将二维图像与三维模型对齐, 再通过调整三维模型的三角面片来调整二维图像相应的像素以获取正面图像. Asthana 等 [5] 将非正面人脸图像映射到一个对齐的三维人脸模型上, 通过调整这个三维模型的姿态获得正面人脸图像. Hassner 等 [6] 使用一个平均三维表观模型近似所有输入人脸图像的形状, 提高重构三维模型的准确率. Fang 等 [7] 基于平均三维模型构建形变模型以重构人脸图像的三维形状和纹理特征. 基于三维模型的方法虽然可以从多姿态人脸图像合成正面人脸图像, 但是三维数据量过大, 训练和优化过程十分缓慢, 且三维数据不易获取.

另一个正面化的直观想法是直接从多姿态人脸图像恢复正面人脸图像. Prince 等 [8] 提出正面图像通过一个姿态线性变换可以生成非正面人脸图像, 故求得该变换的逆变换则可得到正面图像的想法. Chai 等 [9] 学习非正面图像和正面图像间的局部线性变换合成正面图像. 此外, 也可以使用多张多姿态图像合成正面人脸图像. Wang 等 [10] 采用三角剖分结合多个多姿态人脸图像的平均特征得到最终的融合正面人脸图像. Li 等 [11] 对多个多姿态人脸图像的像素加权以合成正面人脸图像. Yi 等 [12] 基于分段放射变换和 Poisson 融合方法, 使用多个姿态图像合成正面人脸图像. 这些方法虽然不涉及三维模型, 没有数据致密计算量大的问题, 但是多姿态图像与正面图像间的映射较为复杂难以学习全面, 细节缺失不好把控. 而使用多张图像的方法虽然能够弥补细节, 但是至少一个近正面图像的要求使得该类方法适用范围较窄.

最近, 基于深度网络合成正面人脸图像的方法也逐渐出现. Zhu 等 [2] 提出一个多姿态感知器, 从输入图像中提取身份和姿态特征合成目标图像. Kan 等 [13] 使用堆叠步进自编码器学习多姿态图像与正面人脸图像间的复杂非线性变换, 将较大姿态图像分多步映射到正面姿态, 但是细节缺失过大导致识别率低. Ouyang 等 [14] 在 Kan [13] 的模型基础上引入非负约束稀疏自编码器学习人脸细节信息, 但是合成图像的效果并没有明显改善且图像尺寸过小实际效用过低. Yim 等 [15] 提出一种多任务卷积神经网络模型, 通过辅助网络重构输入图像和姿态编码, 可以合成保留身份的多姿态图像. Ghodrati 等 [16] 通过组合模型改善合成图像效果, 但是模型参数过多且非端到端模型不方便实际应用. 基于生成式对抗网络 (generative adversary network, GAN) 的模型 [17] 可以生成大角度姿态的正面图像, 但是需要多个 GAN 网络协作以及标注精准的面部关键点辅助, 对硬件设备和图像预处理的要求过高. Tran 等 [18] 根据多个输入图像生成指定姿态图像, 但同样需要多个网络协作. 基于 GAN 的方法需要训练大量参

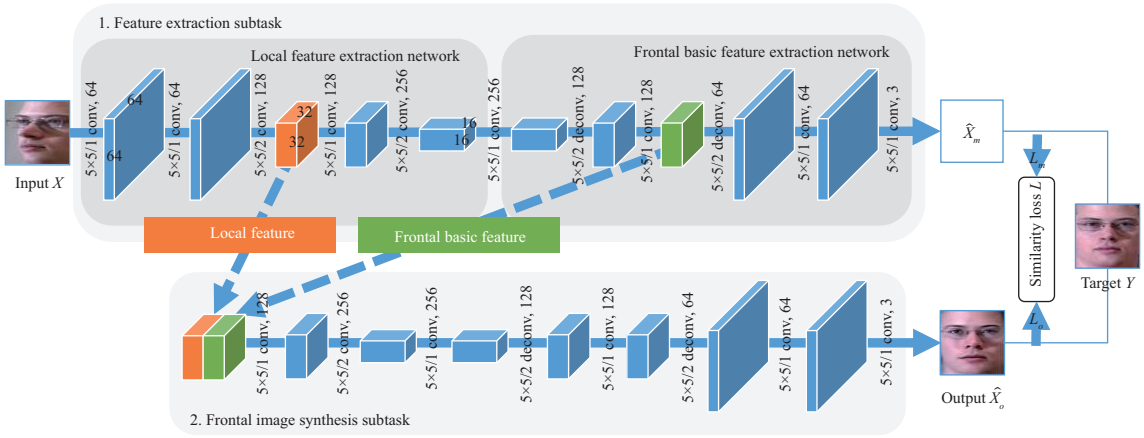

图 1　多姿态人脸图像正面化网络结构图

Figure 1　Multi-pose face frontalization network structure

数, 且 GAN 稳定性差, 固有的模式坍塌问题 (model collapse) 难以解决.

3　基于编解码网络的多姿态人脸图像正面化模型

本文基于编解码网络, 采用多任务学习 (multi-task learning, MTL) 模式, 引入正面基础特征网络合成正面人脸基础特征, 通过融合多姿态图像的局部特征, 合成质量较高的正面人脸图像.

3.1　模型框架

在单任务学习模式下, 编码网络负责提取输入图像特征, 解码网络负责合成正面人脸图像, 此时解码网络完成的工作有两项: 姿态转正和图像合成. 相较其他任务中的解码网络, 例如图像压缩任务中的解码网络 [19], 此时单一任务模式中解码网络的任务增多了, 需要学习的特征随之增多. 为了学习姿态矫正, 解码网络不得已舍弃了部分局部特征, 虽然可以合成正面人脸图像但是细节缺失严重.

为了改善合成图像的细节效果, 直观的想法是特征优化. 对于细节复杂的人脸图像, 本文采用多任务学习模式, 通过合并多个任务中的样本提高模型泛化能力. 多个任务构成的模型拥有两种相关的参数: 具体子任务的参数和所有任务共享的通用参数. 具体子任务的参数只能从各子任务的样本中得到优化; 所有任务共享的通用参数可以从所有任务的汇集数据中得到优化. 共享参数的样本数量相对单任务增加的更多, 可以有效改善泛化性能. 采用 MTL 方式构建模型, 利用多任务间的促进、制约关系促使模型可以从多方向进行优化, 更好地提高模型性能 [20].

多任务学习模式将正面化任务分解为特征解析和图像合成两个子任务. 特征解析子任务包括局部特征解析和正面基础特征解析两个模块; 图像合成子任务包括正面人脸图像合成模块, 如图 1 所示. 为了合成细节更丰富的正面人脸图像, 本文引入正面基础特征网络, 该网络用于合成正面人脸基础特征. 这一特征虽然在人脸细节上有所缺失, 但是可以作为改进的基础. 如此, 原本从多姿态人脸图像直接合成正面人脸图像的任务被改变, 新任务为以多姿态人脸局部特征优化正面人脸基础特征的细节, 从而合成更加细致的正面人脸图像.

图 1 中, 在特征解析子任务中输入尺寸为 64×64×3 的多姿态人脸图像 X, 局部特征解析网络和正面基础特征网络分别从中获取多姿态人脸局部特征 (橙色) 和正面基础特征 (绿色). 两种特征在图

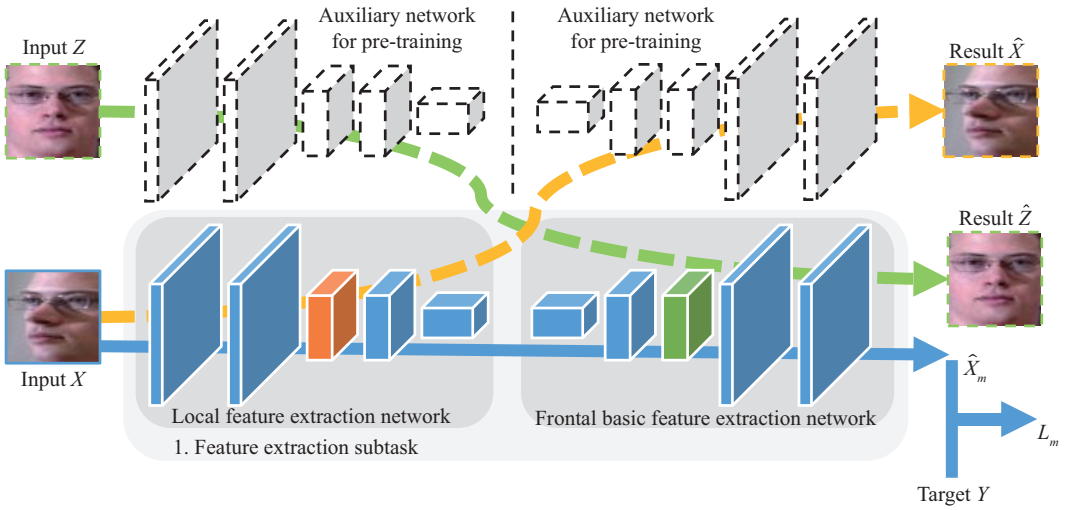

图 2 特征解析子任务网络结构图

Figure 2 Feature analysis subtask network structure

像合成子任务中堆叠在一起, 最终输出正面人脸图像 $\hat{X}_o$. 为了控制正面基础特征网络获取正面姿态特征, 中间结果 $\hat{X}_m$ 与目标图像 Y 计算相似度损失 L_m. $\hat{X}_m$ 为 64×64×3 大小的特征数据, 可视化后亦为正面人脸图像, 实验结果表明该图像在清晰度上比 $\hat{X}_o$ 稍差. 同理, $\hat{X}_o$ 与目标图像 Y 计算相似度损失 L_o. L_m 与 L_o 通过加权求和得到总相似度损失 L, 通过最小化 L 优化网络参数.

3.2 特征解析

特征解析子任务包括局部特征解析网络和正面基础特征网络两个模块. 为了提升合成图像的细节效果, 本文提出在正面基础特征上进行细节补偿的方法. 由于本文模型的输入为多姿态图像, 所以正面基础特征需要根据提取的多姿态图像特征进行合成. 图 2 中虚线部分仅用于辅助局部特征解析网络和正面基础特征网络进行预训练, 模型整体训练时直接将经过预训练的局部特征解析网络与正面基础特征网络连接.

(1) 局部特征解析. 图 2 中, 模型预训练阶段按照黄色虚线箭头流程构建模型, 局部特征解析网络与黑色虚线解码网络构成自编码器, 输出 $\hat{X}$ 尽可能接近输入 X, 输入为多姿态人脸图像, 损失函数为 $L(X,\hat{X})$. 在 MCEDN 模型中使用该预训练网络提取多姿态人脸图像的局部特征, 用以合成正面人脸基础特征和补充人脸细节.

(2) 正面基础特征解析. 为了获取正面人脸特征, 采用深度解码网络构成正面基础特征网络, 通过 MCEDN 模型中编码网络 (局部特征解析网络) 提取的多姿态图像特征合成正面人脸基础特征. 图 2 中, 在模型预训练阶段按照绿色虚线箭头流程构建模型, 黑色虚线编码网络与正面基础特征网络构成自编码器, 输出 $\hat{Z}$ 尽可能接近输入 Z, 输入为正面人脸图像, 损失函数为 $L(Z,\hat{Z})$. 在 MCEDN 模型中使用该预训练网络进行多任务联合训练, 合成的正面基础特征在损失函数 $L_m = L(Y,\hat{X}_m)$ 的约束下完成姿态矫正的任务, 合成的人脸特征也会尽量向 $L(Z,\hat{Z})$ 约束下合成的正面人脸特征靠近, 但是相比后者还是无法避免损失部分细节特征, 所以用此特征直接合成的正面人脸图像的细节效果稍差, 但是其可以作为待改进的基础特征.

可以考虑首先使用基本卷积编解码网络根据输入的多姿态图像合成较为粗糙的正面人脸图像, 然

后再使用一个双输入卷积编解码网络从该合成图像和多姿态图像中分别提取正面基础特征和多姿态局部特征, 之后融合两种特征以强化细节从而合成更清晰的正面人脸图像. 与本文方法相比, 这种双阶段卷积编解码网络的思想更容易理解, 但是如此不仅会增加模型的深度, 还会因为 "从合成的粗糙正面人脸图像提取特征" 的过程造成特征的再次流失, 所以直接在正面基础特征上进行改进.

特征解析子任务的算法伪代码实现见算法 1.

Algorithm 1 Feature synthesis

Require: D: multi-pose face image set; R: frontal face image set; B: batch size; T: number of updates; η: learning rate; θ: trainable parameter set for this subtask; $\theta_{i,j}$: trainable parameter set for the i-th to j-th layers of the subtask; α: weights of the similarity loss of the subtask; β: weights of the similarity loss of the image synthesis task.

1: **for** $t = 0, \ldots, T$ **do**
2: The B images are sampled from training set D as the current training data set X. The B frontal images in the data set R corresponding to the images in the data set X are taken as the current target data set Y;
3: $F_l \leftarrow f_{\theta_{1,3}}(X)$; // F_l: local features of multi-pose face images
4: $F_g \leftarrow f_{\theta_{4,8}}(F_l)$; // F_g: global features of frontal face images
5: $\hat{X}_m \leftarrow f_{\theta_{9,11}}(F_g)$; // $\hat{X}_m$: output of the subtask
6: $L_m \leftarrow \frac{1}{B}\sum_{i=1}^{B} \|\hat{X}_m^i - Y^i\|_2^2$; // L_m: similarity loss of feature synthesis task
7: $L \leftarrow \alpha L_m + \beta L_o$; // L_o: similarity loss of the image synthesis task; L: similarity loss of MCEDN
8: $\theta \leftarrow \mathrm{Adam}(\theta; L; \eta)$;
9: **end for**

3.3 正面人脸图像合成

图像合成子任务只有正面图像合成网络, 该网络与特征解析子任务中的局部特征解析网络可以构成单任务模式下的正面化模型. 由于正面基础特征网络合成了正面基础特征, 完成了姿态正面化的任务, 所以, 多任务模式下的正面图像合成网络只需要专注于更好地强化正面人脸图像细节的问题.

本文模型基于编解码网络架构, 而自编码器是编解码网络的典型应用, 为追求 "更小的编码长度, 更主要的特征信息" 的目标, 自编码器多使用池化层筛选出主要特征, 在图像压缩任务中甚至还要对编码长度进行特别约束. 在这些设置的引导下, 自编码器会抛弃更多的小特征, 而这些小特征往往都是细节特征. 如 Mayya 等 [21] 的文献所示, 池化层输出的特征图比卷积层输出的特征图更加模糊, 损失了细节特征, 所以池化降维法对于图像合成任务有极大的消极影响. 为了提升合成图像的效果, 本文使用卷积层替代池化层. 类似地, 为了从源头降低特征损失, 在特征解析子任务中同样使用这一策略. 卷积层使用不同的步长可以控制生成特征图 (feature map) 的尺寸, 通过卷积核 (convolution kernel) 的区域范围加权可以在考虑邻域特征的情况下学习到更好的特征. 卷积层的这两个属性有利于合成更精细和更准确的图像. 图像合成子任务的算法伪代码实现见算法 2.

3.4 模型实现

MCEDN 基于编解码网络主要由对称的卷积层和转置卷积层构成, 此外还有融合特征所用的联合层. 除联合层外, 每层之后都使用 ReLU 激活函数 [22].

卷积层和转置卷积层类似, 可以表示为

$$\Phi(F_{\mathrm{in}}) = W_k \odot F_{\mathrm{in}} + B_k, \tag{1}$$

Algorithm 2 Image synthesis

Require: D: multi-pose face image set; R: frontal face image set; B: batch size; T: number of updates; η: learning rate; F_l: local features of multi-pose face images, size is $B \times H \times W \times C_l$; F_g: global features of frontal face images, size is $B \times H \times W \times C_g$; φ: trainable parameter set for this subtask; α: weights of the similarity loss of feature synthesis task; β: weights of the similarity loss of the image synthesis task.

1: **for** $t = 0, \ldots, T$ **do**
2: The B images are sampled from training set D as the current training data set X. The B frontal images in the data set R corresponding to the images in the data set X are taken as the current target data set Y;
3: Get F_l and F_g by Algorithm 1;
4: $F_{\text{concat}} \leftarrow \text{Concat}(F_l, F_g)$; // F_{concat}: size is $B \times H \times W \times (C_l + C_g)$
5: $\hat{X}_o \leftarrow f_\varphi(F_{\text{concat}})$; // $\hat{X}_o$: synthetic frontal face image
6: $L_o \leftarrow \frac{1}{B}\sum_{i=1}^{B} \|\hat{X}_o^i - Y^i\|_2^2$; // L_o: similarity loss of image synthesis task
7: $L \leftarrow \alpha L_m + \beta L_o$; // L_o: similarity loss of the feature synthesis task; L: similarity loss of MCEDN
8: $\varphi \leftarrow \text{Adam}(\varphi; L; \eta)$;
9: **end for**

其中 F_{in} 表示本层输入特征, W_k 表示本层的权重参数, B_k 表示本层的偏置参数, $\odot$ 表示卷积操作或者转置卷积操作. 联合层将多姿态人脸图像局部特征和正面基础特征在特征图数量维度上进行堆叠, 联合层表示为

$$\Psi(F_{n\times h\times w\times c_1}, \hat{F}_{n\times h\times w\times c_1}) = F'_{n\times h\times w\times (c_1+c_2)}, \tag{2}$$

其中 n 表示批处理数据大小, h, w 表示特征图的尺寸, c_1 和 c_2 分别表示局部特征 F 和正面基础特征 $\hat{F}$ 的特征图数量, F' 表示堆叠后的特征.

实验使用图像数据对 $\{X^i, Y^i\}$ 训练网络模型. X^i 表示第 i 个多姿态图像, 即输入图像; Y^i 表示与 X^i 对应的正面图像, 即目标图像.

特征解析子任务输出的中间结果 $\hat{X}_m$ 尽可能与目标图像 Y 近似, 图像合成子任务输出的合成图像 $\hat{X}_o$ 与 Y 尽量靠近, 相似程度可以采用相似误差衡量. 两个子任务的相似误差通过均方误差 (mean square error, MSE) 建立, 如下式所示:

$$L_m = \frac{1}{N}\sum_{i=1}^{N} \|G_m(X^i, W) - Y^i\|_2^2, \tag{3}$$

$$L_o = \frac{1}{N}\sum_{i=1}^{N} \|G_o(X^i, W) - Y^i\|_2^2. \tag{4}$$

如此, 整个网络的损失函数如下式所示:

$$L_{\text{total}} = \alpha L_m + \beta L_o, \tag{5}$$

其中, L_m 表示特征解析子任务的相似性损失, L_o 表示图像合成子任务的相似性损失, α 和 β 分别表示两个子任务损失 L_m 和 L_o 的权重系数. 假设两个任务的重要性相同, 则在之后的实验中 α 和 β 统一设置为 1.

4 实验

本节从多方面对基于编解码网络的多姿态人脸图像正面化方法 (MCEDN) 合成的图像进行了效果评估, 并使用合成图像进行了人脸分析实验. 实验在 Tensorflow 平台上进行, 模型使用 Python 语言

实现.

4.1 实验数据集

Multi-PIE 数据集包含 755370 个 RGB 图像, 图像尺寸为 640×480. 337 名被试中有 235 名男性, 102 名女性. 15 个相机从不同角度采集图像, 其中在水平面上有 13 个间隔为 15° 的相机, 侧上方另有 2 个相机. 采集过程中, 通过控制闪光灯产生了 19 种不同的光照条件. 在一些录制过程中, 被试人员按照要求表演微笑、尖叫、蔑视、厌恶、惊讶 5 种表情.

CAS-PEAL-R1 大规模中国人脸图像数据集, 包括 1040 名被试的 30900 个 360×480 的灰度图像. 其中姿态子库包含 1040 名被试的 21840 个图像, 每名被试包括 3 种不同俯仰角情况下水平等距离的 7 种角度变化共 21 个姿态图像.

LFW 是常用于无约束人脸识别任务的数据集, 包括 5729 个不同人的不同表情、姿态、遮挡特征的 13223 个图像. CelebA 大规模人脸属性数据集包括 10177 位名人的 202599 个图像, 每个图像有 40 个属性标签和 5 个关键点, 涉及各种姿态和复杂背景, 可以用于人脸属性识别、人脸检测等研究.

本节在 Multi-PIE [23], CAS-PEAL-R1 [24] 多姿态人脸图像数据集上训练模型. 为了体现模型的泛化性能, 在 LFW [25], CelebA [26] 数据集上进行效果测试, 并与 HPEN [4] 和 Hassner 等 [6] 的方法进行对比.

4.2 模型参数与训练

本文使用的网络参数设置如表 1 所示. 图 1 中模型的两个子任务的相似部分的参数设置相同, 此处仅给出特征解析子任务参数. 参数设置对于不同的实验数据可以灵活改变. 本文实现了 4 种不同模型: 基本卷积编解码网络, 结构如表 1 所示; 3.2 小节所提到的双阶段卷积编解码网络, 该网络在基本卷积编解码网络基础上又添加一个双输入基本卷积编解码网络; MCEDN 和 MCEDN (迁移训练), 结构如图 1 所示. 模型全部使用 Multi-PIE 数据集训练, 迁移训练在 CAS-PEAL-R1 数据集上进行. 在 Multi-PIE 数据集中随机抽取 150 名被试的数据进行训练, 其余数据用于测试. 训练数据包括 19 种不同光照和 7 种不同姿态 ($-45° \sim +45°$). 测试数据包括 19 种不同光照和 6 种不同姿态 (除了正面姿态). 在 CAS-PEAL-R1 数据集中, 随机抽取 500 名被试的数据作为训练集, 包括水平角度上的 7 种姿态, 共 3500 个图像; 其余 540 名被试的数据作为测试集, 包括 3780 个图像. 4 种模型训练效率如表 2 所示.

4.3 实验结果与分析

4.3.1 合成图像效果评估

(1) 合成效果评价. 表 3 和 4 使用结构相似性 (structural similarity, SSIM) [27] 和峰值信噪比 (peak signal to noise ratio, PSNR) 衡量正面合成图像与目标图像的相似度. SSIM 从亮度、对比度、结构 3 方面进行度量, 可以较好地反映人眼主观感受. PSNR 基于 MSE 进行定义, 广泛用于图像评估. SSIM 和 PSNR 的结果数值越大, 图像的失真越小.

基本卷积编解码网络、双阶段卷积编解码网络、MCEDN 的合成图像与目标图像的相似度量化结果如表 3 所示. 从表中可以看出, 双阶段卷积编解码网络的效果比基本卷积编解码网络的效果好且提升幅度很大, 但是 MCEDN 的合成效果更好. 由于双阶段卷积编解码网络相当于两个基本卷积编解码网络拼接在一起, 模型深度大, 需要预先分别训练好两个阶段的网络, 然后再联合训练, 如此模型的训

表 1 网络参数

Table 1 Network parameters

Layers	Input size	Weight/stride	Output size
Conv0	64×64×3	5×5/1	64×64×64
Conv1	64×64×64	5×5/1	64×64×64
Conv2	64×64×64	5×5/2	32×32×128
Conv3	32×32×128	5×5/1	32×32×128
Conv4	32×32×128	5×5/2	16×16×256
Conv5	16×16×256	5×5/1	16×16×256
Deconv6	16×16×256	5×5/2	32×32×128
Conv7	32×32×128	5×5/1	32×32×128
Deconv8	32×32×128	5×5/2	64×64×64
Conv9	64×64×64	5×5/1	64×64×64
Conv10	64×64×64	5×5/1	64×64×3

表 2 不同模型的训练用时

Table 2 Training time for different models

Model	Dataset	Training time (h)
Basic convolutional encoder-decoder network (BCEDN)	Multi-PIE	23
Two-stage convolutional encoder-decoder network (TCEDN)	Multi-PIE	54
MCEDN	Multi-PIE	48
MCEDN (transfer training)	CAS-PEAL-R1	1.5

表 3 3 种模型效果与目标图像的相似度度量结果

Table 3 Similarity evaluation between the results of three models and targets

Method	±45°		±30°		±15°	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
BCEDN	0.5964	16.9613	0.6278	17.3570	0.7053	19.6656
TCEDN	0.7037	19.1310	0.7303	20.1150	0.8023	22.2932
MCEDN	0.7690	23.1557	0.7917	23.4195	0.8719	26.9208

练过程较为复杂不如训练 MCEDN 简单. 实验中, MCEDN 包括两个任务, 任务 1 为特征解析子任务, 其结果为正面基础特征网络的合成图像, 即 $\hat{X}_m$ 可视化. 任务 2 为图像合成子任务, 其结果为合成图像 $\hat{X}_o$, 即 MCEDN 的最终合成结果. 从图 3 中可知, 虽然任务 1 也可以合成正面人脸图像, 但是图像整体效果较为模糊, 细节有所缺失, 例如第 1 列中的眼镜、第 2 列中的胡子. 任务 2 基于合成的正面人脸基础特征, 融合了多姿态图像局部特征, 输出结果在清晰度、细节还原度上都比任务 1 的更好. 表 4 中量化结果也表明任务 2 合成图像的效果优于任务 1 的合成效果. 因此, 引入正面基础特征网络可以有效提升合成图像的效果.

(2) 不同角度的合成图像. 图 4 所示为在 Multi-PIE 数据集上, 6 种姿态的正面合成图像与目标图像的效果对比. 从对比图中可以看出 MCEDN 可以合成不同姿态的正面人脸图像, 而且合成的人脸图像结构稳定, 不同姿态的合成结果没有较大差异.

图 3　每个子任务在 Multi-PIE 数据集上合成的正面人脸图像

Figure 3　Synthesis results by each subtask on Multi-PIE

表 4　任务 1、任务 2 与目标图像的相似度度量结果

Table 4　Similarity evaluation between task1, task2 and targets

Method	±45°		±30°		±15°	
	SSIM	PSNR	SSIM	PSNR	SSIM	PSNR
Task1	0.7511	22.4044	0.7615	21.9180	0.8021	23.0737
Task2	0.7690	23.1557	0.7917	23.4195	0.8719	26.9208

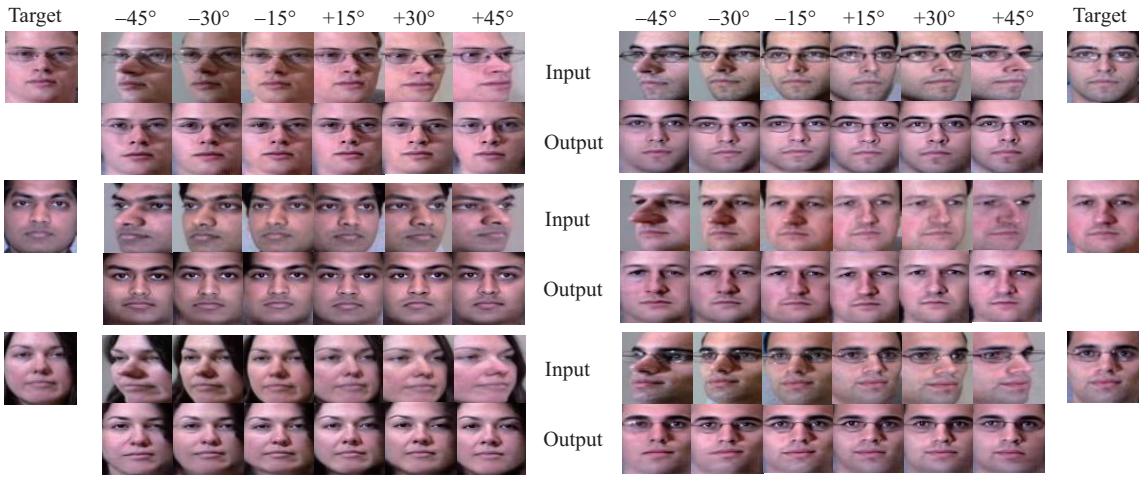

图 4　从不同姿态合成正面人脸图像

Figure 4　Synthesis results by MCEDN under different poses

由于 CAS-PEAL-R1 数据集中数据较少, 为了充分训练模型, 本文使用在 Multi-PIE 数据集上训练的模型作为预训练模型, 然后在 CAS-PEAL-R1 数据集上进行迁移训练, 合成效果如图 5 所示. 从图 5 可知, MCEDN 模型使用不同的训练数据集都能达到很好的合成效果.

(3) 不同数据集上的合成图像. 为了测试模型的泛化性能, 在 LFW, CelebA 数据集上测试仅使用 Multi-PIE 数据集训练的模型, 合成图像效果如图 6 所示. 从图 6 中可知, 虽然 MCEDN 模型合成的图像在色调上偏向于 Multi-PIE 数据集, 但是其合成的图像相较其他两种方法更加自然且保留了更多的人脸身份特征和表情特征, 具有良好的泛化性能. 从第 3 和 4 列可以看出, MCEDN 模型对垂直

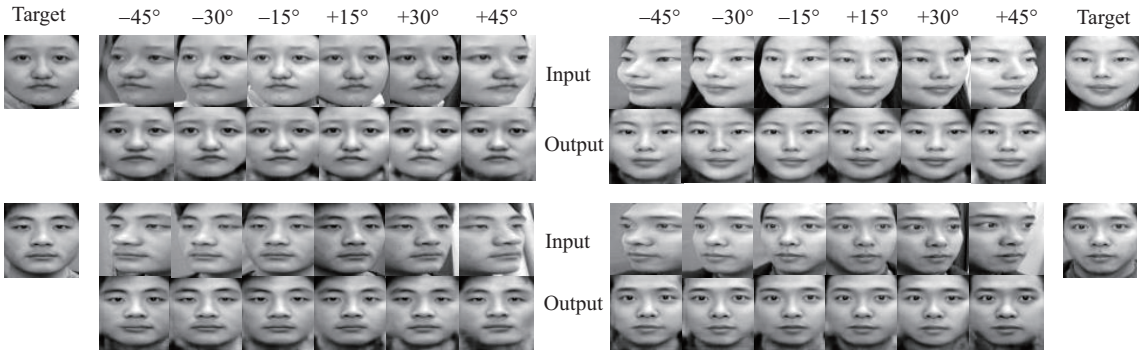

图 5 在 CAS-PEAL-R1 数据集上合成不同姿态的正面人脸图像

Figure 5 Synthesis results by MCEDN under different poses on CAS-PEAL-R1 dataset

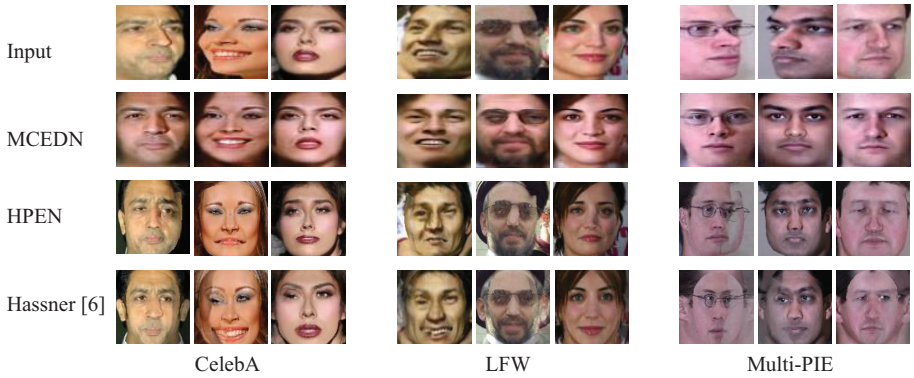

图 6 MCEDN 与其他方法在 3 种不同数据集上的正面人脸合成图像结果

Figure 6 Synthesis results on different datasets by different method

图 7 不同光照条件下合成效果展示

Figure 7 Synthesis results under various illuminations

角度姿态变化具有鲁棒性, 即使人脸图像有垂直角度变化, 也可以完成水平方向上的正面化. HPEN[4] 虽然可以合成较大角度姿态的正面图像, 但是无法保留细节特征, 例如第 2 和 4 列的输入图像中微笑的表情在正面化图像中消失; 在第 7 列的大角度姿态情况下, 身份特征都遭到了一定程度的损坏且引入了大量噪声. Hassner 等 [6] 基于人脸对称性补全合成图像, 但是在稍大角度姿态情况下, 补全的虚影已经严重影响了视觉效果.

(4) 不同光照条件下的合成图像. 图 7 为在 Multi-PIE 数据集上合成的不同光照条件的正面图像. 对比图 7 中图像可知, 在相似性损失约束下, MCEDN 模型对于光照变化具有一定鲁棒性. 在不同的光照条件下, MCEDN 合成的图像全局结构稳定, 保留了局部细节, 人脸肤色与输入图像一致.

表 5 表情识别结果

Table 5 Rank-1 face expression recognition rate

Method	±45°	±30°	±15°
Original images	0.6503	0.8557	0.9484
Ref. [15]	0.8511	0.9127	0.9515
Task1	0.8687	0.9285	0.9501
Task2	**0.9224**	**0.9439**	**0.9618**

表 6 人脸识别结果

Table 6 Rank-1 face recognition rate

Method	±45°	±30°	±15°
Ref. [15]	0.8838	0.9526	0.9714
Task1	0.8918	0.9584	0.9746
Task2	**0.9136**	**0.9615**	**0.9803**

4.3.2 人脸分析实验

为量化展示模型保留图像细节特征的能力, 采用 Ding 等 [28] 的卷积神经网络模型在 Multi-PIE 数据集上进行表情识别, 训练样本包括中性、微笑、惊讶、蔑视、厌恶、尖叫 6 种表情, 跨越 19 种光照, $-45° \sim +45°$ 间 7 种姿态变化的图像; 测试集不包括正面图像, 与训练集的挑选方式相同; 采用 Light CNN [29] 在 Multi-PIE 数据集上进行人脸识别, 数据包括 6 种表情、正面光照、7 种姿态变化, 训练集和测试集的抽取方式与训练 MCEDN 模型的相同. 表情识别结果如表 5 所示, 人脸识别结果如表 6 所示. 其中, 任务 1 为正面基础特征网络的合成图像 $\hat{X}_m$ 识别结果, 任务 2 为图像合成子任务的合成图像 $\hat{X}_o$ 亦即 MCEDN 的合成图像的识别结果. 文献 [16] 通过辅助网络提升合成图像的质量, 本小节使用该方法进行对比实验.

从表 5 和 6 可知, MCEDN 模型在两个识别任务的所有角度上都取得了最高的识别准确率. 人脸姿态的角度越大, 合成的正面图像对于识别准确率的提升越大. MTL 模式的运用十分成功, 两个任务在共享参数的协调下都达到了最优; 正面基础特征网络对于正面人脸图像的合成起到了积极作用, 不仅保留了人脸身份特征, 还丰富了表情细节特征. 在大角度姿态情况下, 任务 2 较任务 1 提升了 6% 的表情识别率, 较文献 [15] 的方法提升了 7% 的表情识别率.

综上所述, MCEDN 模型可以在尽可能多的保留输入图像细节特征的情况下, 合成整体结构稳定的正面人脸图像, 尤其是在保留人脸身份特征和表情特征方面, 这使得该模型可以用于人脸识别和表情识别任务. 此外, MCEDN 模型对于不同数据集、垂直角度变化都具有良好的泛化性能, 在自然条件数据集 (CelebA, LFW) 上都可以取得很好的合成效果.

5 结束语

本文提出了一种基于编解码网络的多姿态人脸图像正面化方法 —— MCEDN 模型, 该模型采用了全卷积网络, 保留了更多像素间的关联特征, 对于从多姿态图像合成正面人脸图像的任务有良好的实验效果. 该模型通过引入正面基础特征网络合成正面人脸基础特征, 然后在此基础上融合多姿态图像的局部特征进行细节补偿, 最终合成更为清晰的正面人脸图像. 模型包括两个子任务, 两个任务共享

部分参数. 特征解析子任务使用编码网络提取多姿态人脸图像的局部特征、使用正面基础特征网络合成正面人脸基础特征; 图像合成子任务使用解码网络融合两种特征合成正面人脸图像. 实验结果表明, MCEDN 具有很好的泛化能力, 在多个数据集上都可以合成结构稳定、细节清晰的正面人脸图像. 模型保留了表情特征和身份特征, 合成的人脸图像可以用于表情识别和人脸识别任务.

目前, 本文提出的模型对于合成其他数据集的正面图像还存在色度偏向训练数据集的问题. 在未来的工作中, 我们将考虑保留光照特征并引入越层连接 (skip connections) 保留更多的细节特征, 合成效果更好的正面图像.

参考文献

1 Zhu Z Y, Luo P, Wang X G, et al. Deep learning identity-preserving face space. In: Proceedings of the IEEE International Conference on Computer Vision, Sydney, 2013. 113–120

2 Zhu Z Y, Luo P, Wang X G, et al. Multi-view perceptron: a deep model for learning face identity and view representations. In: Proceedings of the Advances in Neural Information Processing Systems, Montreal, 2014. 217–225

3 Argyriou A, Evgeniou T, Pontil M. Multi-task feature learning. In: Proceedings of the 20th Annual Conference on Neural Information Processing Systems, Vancouver, 2006. 41–48

4 Zhu X Y, Lei Z, Yan J J, et al. High-fidelity pose and expression normalization for face recognition in the wild. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 787–796

5 Asthana A, Marks T K, Jones M J, et al. Fully automatic pose-invariant face recognition via 3D pose normalization. In: Proceedings of the International Conference on Computer Vision, Barcelona, 2011. 937–944

6 Hassner T, Harel S, Paz E, et al. Effective face frontalization in unconstrained images. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 4295–4304

7 Fang S Y, Zhou D K, Cao Y P, et al. Frontal face image synthesis based on pose estimation. Comput Eng, 2015, 41: 240–244 [方三勇, 周大可, 曹元鹏, 等. 基于姿态估计的正面人脸图像合成. 计算机工程, 2015, 41: 240–244]

8 Prince S J D, Warrell J, Elder J H, et al. Tied factor analysis for face recognition across large pose differences. IEEE Trans Pattern Anal Mach Intel, 2008, 30: 970–984

9 Chai X J, Shan S G, Chen X L, et al. Locally linear regression for pose-invariant face recognition. IEEE Trans Image Process, 2007, 16: 1716–1725

10 Wang Y N, Su J B. Multipose face image recognition based on image synthesis. Pattern Recogn Artif Intel, 2015, 28: 848–856 [王亚南, 苏剑波. 基于图像合成的多姿态人脸图像识别方法. 模式识别与人工智能, 2015, 28: 848–856]

11 Li Y L, Feng J F. Multi-view face synthesis using minimum bending deformation. J Comput-Aided Design Comput Graph, 2011, 23: 1085–1090 [李月龙, 封举富. 基于最小扭曲变换的正面人脸图像合成. 计算机辅助设计与图形学学报, 2011, 23: 1085–1090]

12 Yi X B, Chen Y. Frontal face synthesizing based on poisson image fusion under piecewise affine warp. Comput Eng Appl, 2016, 52: 172–177 [仪晓斌, 陈莹. 分段仿射变换下基于泊松融合的正面人脸合成. 计算机工程与应用, 2016, 52: 172–177]

13 Kan M N, Shan S G, Chang H, et al. Stacked progressive auto-encoders (spae) for face recognition across poses. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Columbus, 2014. 1883–1890

14 Ouyang N, Ma Y T, Lin L P. Multi-pose face reconstruction and recognition based on multi-task learning. J Comput Appl, 2016, 37: 896–900 [欧阳宁, 马玉涛, 林乐平. 基于多任务学习的多姿态人脸重建与识别. 计算机应用, 2016, 37: 896–900]

15 Yim J, Jung H, Yoo B, et al. Rotating your face using multi-task deep neural network. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, Boston, 2015. 676–684

16 Ghodrati A, Jia X, Pedersoli M, et al. Towards automatic image editing: learning to see another you. 2015. ArXiv:151108446

17 Huang R, Zhang S, Li T Y, et al. Beyond face rotation: global and local perception gan for photorealistic and identity preserving frontal view synthesis. 2017. ArXiv:170404086

18 Tran L, Yin X, Liu X M. Disentangled representation learning gan for pose-invariant face recognition. In: Proceedings of the Computer Vision and Pattern Recognition, Honolulu, 2017. 1283–1292

19 Theis L, Shi W, Cunningham A, et al. Lossy image compression with compressive autoencoders. 2017. ArXiv: 170300395
20 Goodfellow I, Bengio Y, Courville A, et al. Deep Learning. Cambridge: MIT Press, 2016
21 Mayya V, Pai R M, Pai M M. Automatic facial expression recognition using DCNN. Procedia Comput Sci, 2016, 93: 453–461
22 Nair V, Hinton G E. Rectified linear units improve restricted boltzmann machines. In: Proceedings of International Conference on Machine Learning, Haifa, 2010. 807–814
23 Gross R, Matthews I, Cohn J, et al. Multi-PIE. Image Vision Comput, 2010, 28: 807–813
24 Gao W, Cao B, Shan S G, et al. The CAS-PEAL large-scale Chinese face database and baseline evaluations. IEEE Trans Syst Man Cybern A, 2008, 38: 149–161
25 Huang G B, Ramesh M, Berg T, et al. Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments. Technical Report 07-49. 2007
26 Liu Z, Luo P, Wang X G, et al. Deep learning face attributes in the wild. In: Proceedings of the IEEE International Conference on Computer Vision, Santiago, 2015. 3730–3738
27 Wang Z, Bovik A C, Sheikh H R, et al. Image quality assessment: from error visibility to structural similarity. IEEE Trans Image Process, 2004, 13: 600–612
28 Ding H, Zhou S K, Chellappa R. Facenet2expnet: regularizing a deep face recognition net for expression recognition. In: Proceedings of the IEEE International Conference on Automatic Face and Gesture Recognition, Washington, 2017. 118–126
29 Wu X, He R, Sun Z A, et al. A light CNN for deep face representation with noisy labels. 2015. ArXiv:151102683

A multi-pose face frontalization method based on encoder-decoder network

Haiyue XU[1,2], Naiming YAO[1,2], Xiaolan PENG[1,2], Hui CHEN[1,2*] & Hongan WANG[1,2,3]

1. *Beijing Key Laboratory of Human-Computer Interaction, Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China*;
2. *University of Chinese Academy of Sciences, Beijing* 100049*, China*;
3. *State Key Laboratory of Computer Science, Institute of Software, Chinese Academy of Sciences, Beijing* 100190*, China*
* Corresponding author. E-mail: chenhui@iscas.ac.cn

Abstract Multi-pose face frontalization can alleviate the influence of pose variance on face analysis. The traditional method of synthesizing a frontal face image directly from a multi-pose face image presents a problem in missing face details. To overcome this problem, we propose a face frontalization method based on the encoder-decoder network, namely multitask convolutional encoder-decoder network (MCEDN). The MCEDN introduces a frontal raw feature network to synthesize the global raw features of the frontal face. Then, the network utilizes the decoder to synthesize a clearer frontal face image by fusing local features extracted by the encoder and global raw features. We use a multitask learning mechanism to build an end-to-end model. The method then integrates three modules, namely local feature extraction, global raw feature synthesis, and frontal image synthesis. The model performance was improved by sharing parameters. In comparison with existing methods, MCEDN can synthesize frontal face images with a stable structure and rich details on multiple datasets. Then, we use the synthesized frontal images for face recognition and face expression recognition, and the state-of-the-art results demonstrate that the MCEDN preserves a number of face details.

Keywords face frontalization, convolutional neural network, encoder-decoder network, multitask learning, face recognition, facial expression recognition

Haiyue XU was born in 1992. She received her B.S. degree in electronic and information engineering from Beijing University of Technology, Beijing, in 2015. Currently, she is a master at Institute of Software, Chinese Academy of Sciences, Beijing and the University of Chinese Academy of Sciences, Beijing. Her research interests include human-computer interaction, machines learning, and computer vision.

Hui CHEN was born in 1974. She received her Ph.D. degree in computer science from the Chinese University of Hong Kong, Hong Kong in 2006. Currently, she is a professor at Institute of Software, Chinese Academy of Sciences. Her research interests include human-computer interaction, affective interaction, haptics, and virtual reality.

Naiming YAO was born in 1986. He received his M.S. degree in computer software and theory from Capital Normal University. Currently, he is a Ph.D. candidate at Institute of Software, Chinese Academy of Sciences, Beijing and the University of Chinese Academy of Sciences, Beijing. His research interests include human-computer interaction, affective computing, machines learning, and computer vision.

Hongan WANG was born in 1963. He received his Ph.D. degree in computer application technology from Institute of Software, Chinese Academy of Sciences, Beijing in 1999. Currently, he is a professor at Institute of Software, Chinese Academy of Science. His research interests include real-time intelligence and user interface.