Phát hiện khuôn mặt trên ảnh tĩnh sử dụng MTCNN

Báo cáo đồ án môn học: Xử lý ảnh

Nguyễn Tất Mạnh (0210768)
Lớp 68CS2, Khoa Công nghệ Thông tin, Trường Đại học Xây dựng Hà Nội
Giảng viên hướng dẫn: Thầy Đào Việt Cường

***Abstract*—Phát hiện khuôn mặt (face detection) là bước tiền đề cho nhiều bài toán thị giác máy tính như nhận dạng, theo dõi và phân tích biểu cảm. Trong môi trường thực tế, khuôn mặt thường xuất hiện với nhiều biến thiên (tư thế, ánh sáng, che khuất, kích thước nhỏ), khiến hiệu năng của các phương pháp truyền thống suy giảm. Trong báo cáo này, nhóm xây dựng hệ thống phát hiện khuôn mặt trên ảnh tĩnh dựa trên MTCNN (Multi-task Cascaded Convolutional Networks). Nhóm triển khai *hai nhánh xử lý* để so sánh: RAW (ảnh gốc) và PRE (lọc Gaussian (kernel 3×3) + cân bằng histogram trên kênh độ sáng). Thực nghiệm cho thấy nhánh RAW đạt trung bình $\mathrm{Avg}(N) = 32.80$, $\mathrm{Avg}(\overline{s}) = 0.995$, $\mathrm{Avg}(t) = 2138.2$ ms; trong khi nhánh PRE đạt $\mathrm{Avg}(N) = 27.67$, $\mathrm{Avg}(\overline{s}) = 0.991$, $\mathrm{Avg}(t) = 1981.7$ ms. Kết quả gợi ý rằng tiền xử lý giúp giảm thời gian xử lý nhưng có thể làm giảm số lượng khuôn mặt phát hiện được trên tập ảnh thử nghiệm; nhóm thảo luận nguyên nhân và đề xuất hướng cải tiến theo ngữ cảnh ảnh.**

***Index Terms*—Phát hiện khuôn mặt, MTCNN, học sâu, tiền xử lý ảnh, histogram equalization, landmarks**

I. Giới thiệu

Phát hiện khuôn mặt là bài toán xác định vị trí (hộp bao) và đôi khi là các điểm đặc trưng (landmarks) của khuôn mặt trong ảnh. Đây là bước đầu vào quan trọng trong chuỗi xử lý của nhiều hệ thống như: nhận dạng danh tính (face recognition), xác thực sinh trắc, giám sát an ninh, lọc ảnh chân dung, hay tương tác người–máy. Trong thực tế, ảnh đầu vào thường chịu ảnh hưởng bởi điều kiện chiếu sáng không đồng đều, nhiễu, độ phân giải thấp, hoặc khuôn mặt có kích thước nhỏ và bị che khuất. Vì vậy, các mô hình phát hiện cần đạt được sự cân bằng giữa *độ chính xác* và *tốc độ*.

Các phương pháp truyền thống như Haar Cascade của Viola–Jones có ưu điểm chạy nhanh nhưng kém bền vững trước biến thiên lớn về tư thế và ánh sáng [2]. Sự phát triển của học sâu đã tạo ra nhiều bộ phát hiện mạnh hơn; trong đó MTCNN là một lựa chọn phổ biến nhờ thiết kế cascade coarse-to-fine và dự đoán đồng thời bbox và landmarks [1]. Một vấn đề thực tế khác là chất lượng ảnh: với ảnh tối hoặc tương phản thấp, việc bổ sung tiền xử lý (lọc nhiễu, tăng tương phản) đôi khi giúp cải thiện khả năng phát hiện, nhưng cũng có thể làm biến dạng đặc trưng nếu cấu hình chưa phù hợp [9].

Đóng góp của báo cáo: (1) triển khai hệ thống phát hiện khuôn mặt dựa trên MTCNN; (2) xây dựng pipeline so sánh hai nhánh RAW/PRE và quy tắc chọn nhánh tốt nhất theo số mặt phát hiện, độ tin cậy và thời gian; (3) báo cáo kết quả thực nghiệm và thảo luận định hướng tối ưu.

II. Cơ sở lý thuyết và nghiên cứu liên quan

A. Bài toán phát hiện khuôn mặt

Bài toán phát hiện khuôn mặt có thể xem như một bài toán phát hiện đối tượng (object detection) chuyên biệt với lớp đối tượng là "face". Đầu ra thường gồm tập các hộp bao $B = \{b_i\}$ và điểm tin cậy (score) $S = \{s_i\}$. Trong nhiều hệ thống hiện đại, mô hình còn dự đoán landmarks $L = \{\ell_i\}$ (ví dụ 5 điểm: hai mắt, mũi, hai khoé miệng) phục vụ căn chỉnh khuôn mặt.

B. Từ phương pháp truyền thống đến học sâu

Phương pháp Viola–Jones [2] sử dụng đặc trưng Haar, ảnh tích phân và cascade AdaBoost để đạt tốc độ cao trên CPU. Tuy nhiên, mô hình này thường gặp khó khăn khi khuôn mặt bị quay, thay đổi biểu cảm và điều kiện ánh sáng phức tạp. Các phương pháp học sâu hiện đại (RetinaFace [4], S^3FD [5], DSFD [6], …) cho thấy hiệu năng vượt trội trên các bộ chuẩn như WIDER FACE [3]. MTCNN [1] là một trong các mô hình tiêu biểu theo hướng *cascade* và *đa nhiệm*, phù hợp cho cả phát hiện và căn chỉnh.

C. MTCNN và học đa nhiệm

MTCNN gồm ba mạng con: P-Net (proposal), R-Net (refine) và O-Net (output). Điểm nổi bật là mô hình học đồng thời nhiều nhiệm vụ (multi-task learning) để khai thác tương quan giữa phân loại mặt/không mặt, hồi quy bbox và hồi quy landmarks [1], [8]. Trong thực nghiệm của nhóm, các trọng số của MTCNN được sử dụng ở chế độ *pre-trained* (không huấn luyện lại).

III. Phương pháp đề xuất

A. Tổng quan hệ thống

Hình 1 mô tả pipeline của nhóm. Ảnh đầu vào được xử lý theo hai nhánh: RAW (giữ nguyên) và PRE (tiền xử lý tăng chất lượng ảnh). Mỗi nhánh được đưa qua cùng một detector MTCNN để thu được bbox, landmarks và score. Sau đó hệ thống tính các chỉ số đánh giá và chọn nhánh tốt nhất theo quy tắc ưu tiên: **số mặt** $N \rightarrow$ **độ tin cậy trung bình** $\overline{s} \rightarrow$ **thời gian** t.

B. Chi tiết MTCNN

Với một ảnh đầu vào I, MTCNN xây dựng tháp ảnh (image pyramid) ở nhiều tỉ lệ và quét bằng P-Net để sinh các ứng viên bbox. Các ứng viên được lọc theo ngưỡng và qua *Non-Maximum Suppression (NMS)* để loại bỏ trùng lặp. R-Net tiếp tục tinh chỉnh và loại bỏ false positive; cuối cùng O-Net đưa ra bbox và landmarks chính xác hơn [1].

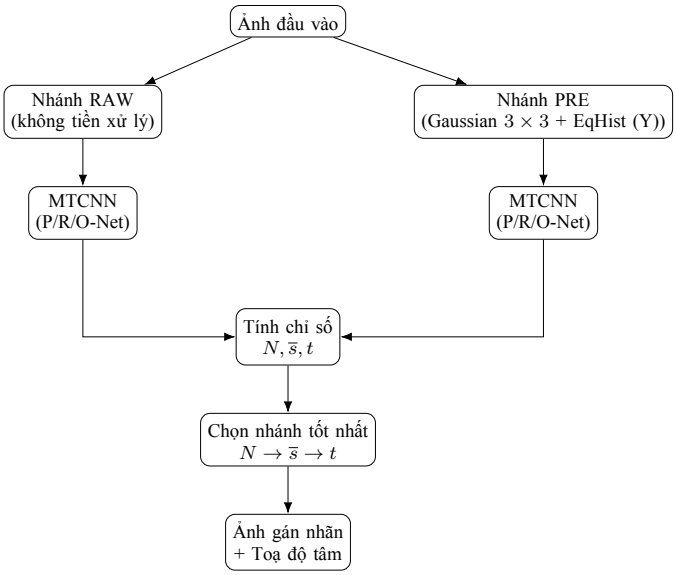


Fig. 1. Pipeline so sánh hai nhánh RAW và PRE.

1) Hồi quy hộp bao: Một bbox được biểu diễn bởi $b = (x_1, y_1, x_2, y_2)$ với $w = x_2 - x_1$ và $h = y_2 - y_1$. Mạng dự đoán vector hiệu chỉnh $\Delta = (\delta_1, \delta_2, \delta_3, \delta_4)$. Bbox sau tinh chỉnh:

$$b' = \big(x_1 + \delta_1 w,\ y_1 + \delta_2 h,\ x_2 + \delta_3 w,\ y_2 + \delta_4 h\big). \quad (1)$$

2) Non-Maximum Suppression và IoU: Độ chồng lấp giữa hai bbox b_a, b_b được đo bởi:

$$\mathrm{IoU}(b_a, b_b) = \frac{\mathrm{area}(b_a \cap b_b)}{\mathrm{area}(b_a \cup b_b)}. \quad (2)$$

NMS giữ lại bbox có score cao hơn và loại các bbox có IoU vượt ngưỡng.

3) Học đa nhiệm và hàm mất mát: MTCNN tối ưu đồng thời ba thành phần mất mát [1], [8]:

$$\mathcal{L}_{cls} = -\big[y \log(p) + (1-y)\log(1-p)\big], \quad (3)$$

$$\mathcal{L}_{box} = \left\| \hat{\Delta} - \Delta \right\|_2^2, \quad (4)$$

$$\mathcal{L}_{lm} = \left\| \hat{\ell} - \ell \right\|_2^2, \quad (5)$$

$$\mathcal{L} = \alpha \mathcal{L}_{cls} + \beta \mathcal{L}_{box} + \gamma \mathcal{L}_{lm}, \quad (6)$$

trong đó p là xác suất "face", $y \in \{0, 1\}$ là nhãn, $\ell \in \mathbb{R}^{10}$ là landmarks (5 điểm 2D), và α, β, γ là trọng số.

C. Tiền xử lý ảnh (nhánh PRE)

Nhánh PRE nhằm giảm nhiễu và tăng tương phản để hỗ trợ phát hiện trong các điều kiện ảnh kém. Nhóm áp dụng hai bước chính:

- **Lọc Gaussian** (kernel 3×3, σ tự động) nhằm làm mượt nhiễu cao tần. Bộ lọc Gaussian 2D:

$$G(x, y) = \frac{1}{2\pi\sigma^2} \exp\left(-\frac{x^2 + y^2}{2\sigma^2}\right), \quad (7)$$

và ảnh sau lọc $I' = I * G$ [9].

TABLE I
Tham số cấu hình chính của hệ thống

Tham số	**Giá trị**
`min_face_size`	20
`steps_threshold` (P/R/O)	(0.6, 0.7, 0.7)
`scale_factor`	0.709
`min_confidence` (lọc kết quả)	0.8
Tiền xử lý PRE	Gaussian blur + EqHist (kênh Y)

- **Cân bằng histogram** trên kênh độ sáng để tăng tương phản toàn cục. Với mức xám rời rạc r_k và hàm phân phối tích luỹ $CDF(r_k)$, phép biến đổi:

$$s_k = (L-1)\, CDF(r_k), \quad (8)$$

trong đó L là số mức xám (thường $L = 256$) [9].

Trong cài đặt, nhóm chuyển ảnh sang không gian màu YCrCb, áp dụng equalization trên kênh Y rồi chuyển ngược về RGB để giảm biến dạng màu so với equalization độc lập từng kênh.

IV. Cài đặt hệ thống

A. Công cụ và thư viện

Hệ thống được cài đặt bằng Python, sử dụng: `TensorFlow` để chạy mô hình MTCNN, `OpenCV` để xử lý ảnh và I/O [10].

B. Tham số thực nghiệm

Bảng I liệt kê các tham số chính được nhóm sử dụng (bám theo cấu hình trong notebook/demo). Ngưỡng tin cậy tối thiểu được đặt ở mức 0.8 để giảm false positive.

C. Xuất kết quả

Với mỗi ảnh đầu vào, hệ thống lưu: (i) ảnh đã vẽ bbox và landmarks, (ii) file toạ độ tâm khuôn mặt. Tâm mỗi bbox $b_i = (x_1, y_1, x_2, y_2)$ được tính:

$$c_i = \left(\frac{x_1 + x_2}{2}, \frac{y_1 + y_2}{2}\right). \quad (9)$$

V. Thực nghiệm và kết quả

A. Thiết lập thực nghiệm

Nhóm tiến hành thử nghiệm trên tập ảnh tĩnh (định dạng `.jpg`/`.png`). Mỗi ảnh được xử lý độc lập qua hai nhánh RAW/PRE, sau đó áp dụng quy tắc chọn nhánh tốt nhất. Lưu ý: do phạm vi đồ án không xây dựng ground-truth nhãn bbox, nhóm tập trung so sánh *tương đối* giữa hai nhánh dựa trên số lượng phát hiện, độ tin cậy trung bình và thời gian xử lý.

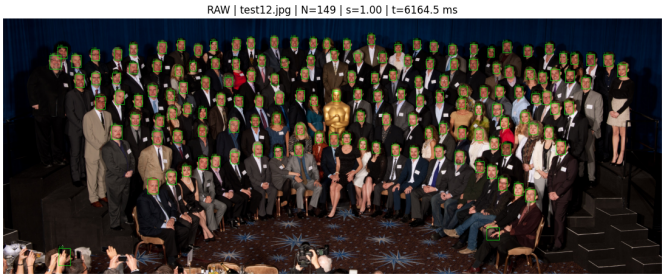
B. Chỉ số đánh giá

Với mỗi ảnh, nhóm đo:

- N: số khuôn mặt phát hiện (sau khi lọc theo `min_confidence`).
- $\bar{s}$: điểm tin cậy trung bình của các khuôn mặt phát hiện.
- t (ms): thời gian xử lý của nhánh (bao gồm tiền xử lý nếu có).

TABLE II
Kết quả trung bình trên tập kiểm thử

Nhánh	Avg(**N**)	Avg($\bar{\mathbf{s}}$)	Avg(**t**) (ms)
RAW	32.80	0.995	2138.2
PRE	27.67	0.991	1981.7



(a) RAW

(b) PRE

Fig. 2. So sánh kết quả phát hiện khuôn mặt giữa RAW và PRE trên cùng một ảnh.

C. Kết quả trung bình

Bảng II tổng hợp kết quả trung bình của nhóm trên tập ảnh thử nghiệm.

D. Nhận xét

So với RAW, nhánh PRE có thời gian xử lý trung bình giảm khoảng 7.3% nhưng số lượng khuôn mặt phát hiện giảm khoảng 15.6%. Một số nguyên nhân có thể gồm: (i) cân bằng histogram toàn cục có thể làm thay đổi phân bố cường độ tại vùng da mặt, ảnh hưởng đến đặc trưng mà mạng đã học; (ii) lọc Gaussian có thể làm mờ biên khi khuôn mặt nhỏ, khiến điểm tin cậy giảm và bị loại bởi ngưỡng 0.8; (iii) tập ảnh thử nghiệm có thể chứa nhiều ảnh đủ sáng, khiến PRE không mang lại lợi ích rõ rệt. Tuy nhiên, trong các ảnh tối hoặc tương phản thấp, PRE có thể giúp phát hiện ổn định hơn (đánh giá định tính), do tăng độ tương phản vùng mặt [9].

VI. Kết luận và hướng phát triển

A. Kết luận

Báo cáo đã trình bày hệ thống phát hiện khuôn mặt trên ảnh tĩnh dựa trên MTCNN và pipeline so sánh hai nhánh RAW/PRE. Kết quả thực nghiệm cho thấy nhánh PRE giúp giảm thời gian xử lý trung bình nhưng làm giảm số lượng khuôn mặt phát hiện được trong tập thử nghiệm của nhóm; theo quy tắc chọn nhánh, RAW thường được ưu tiên khi cần tối đa số lượng khuôn mặt phát hiện.

B. Hướng phát triển

Một số hướng cải tiến:

- Thiết kế cơ chế *chọn tiền xử lý thích ứng* dựa trên độ sáng/độ tương phản của ảnh thay vì áp dụng cố định.
- Bổ sung đánh giá có *ground-truth* (ví dụ dùng WIDER FACE [3]) để đo Precision/Recall/AP.
- Tối ưu tốc độ bằng cách resize ảnh đầu vào, batch inference hoặc sử dụng GPU; hoặc thay thế bằng các detector một giai đoạn như RetinaFace [4].

References

[1] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao, "Joint Face Detection and Alignment Using Multitask Cascaded Convolutional Networks," *IEEE Signal Processing Letters*, vol. 23, no. 10, pp. 1499–1503, 2016, doi: 10.1109/LSP.2016.2603342.

[2] P. Viola and M. J. Jones, "Rapid Object Detection using a Boosted Cascade of Simple Features," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2001, doi: 10.1109/CVPR.2001.990517.

[3] S. Yang, P. Luo, C.-C. Loy, and X. Tang, "WIDER FACE: A Face Detection Benchmark," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2016, pp. 5525–5533.

[4] J. Deng, J. Guo, E. Ververas, I. Kotsia, and S. Zafeiriou, "RetinaFace: Single-Shot Multi-Level Face Localisation in the Wild," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2020, pp. 5203–5212.

[5] S. Zhang, X. Zhu, Z. Lei, H. Shi, X. Wang, and S. Z. Li, "S^3FD: Single Shot Scale-invariant Face Detector," in *Proc. IEEE Int. Conf. Computer Vision (ICCV)*, 2017.

[6] J. Li *et al.*, "DSFD: Dual Shot Face Detector," in *Proc. IEEE/CVF Conf. Computer Vision and Pattern Recognition (CVPR)*, 2019, pp. 5060–5069, doi: 10.1109/CVPR.2019.00520.

[7] X. Zhu and D. Ramanan, "Face detection, pose estimation, and landmark localization in the wild," in *Proc. IEEE Conf. Computer Vision and Pattern Recognition (CVPR)*, 2012, pp. 2879–2886, doi: 10.1109/CVPR.2012.6248014.

[8] R. Caruana, "Multitask Learning," *Machine Learning*, vol. 28, pp. 41–75, 1997, doi: 10.1023/A:1007379606734.

[9] R. C. Gonzalez and R. E. Woods, *Digital Image Processing*, 4th ed. Pearson, 2018.

[10] G. Bradski, "The OpenCV Library," *Dr. Dobb's Journal of Software Tools*, 2000.

[11] G. Liu, J. Xiao, and X. Wang, "Optimization of Face Detection Algorithm based on MTCNN," *International Core Journal of Engineering*, vol. 7, no. 8, pp. 456–464, 2021, doi: 10.6919/ICJE.202108_7(8).0067.

[12] Q. Trần, "Nhận diện khuôn mặt với mạng MTCNN và FaceNet (Phần 1)," Viblo, 2021. [Online]. Available: https://viblo.asia/p/nhan-dien-khuon-mat-voi-mang-mtcnn-va-facenet-phan-1-Qbq5QDN4lD8.