Phát hiện tự động các bộ phận của cây từ ảnh sử dụng mạng nơ-ron tích chập

Nguyễn Thị Thanh Nhàn[1,2], Lê Thị Lan[1], Vũ Hải[1], Hoàng Văn Sâm[3]
[1]Viện nghiên cứu quốc tế MICA, Trường Đại học Bách khoa Hà Nội
[2]Khoa Công nghệ thông tin, Trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên
[3]Bộ môn Thực vật rừng, Trường Đại học Lâm nghiệp

Tác giả liên hệ: Nguyễn Thị Thanh Nhàn, nttnhan@ictu.edu.vn
Ngày nhận bài: 27/11/2017, ngày sửa chữa: 08/05/2018, ngày duyệt đăng: 21/05/2018
Xem sớm trực tuyến: 08/11/2018, định danh DOI: 10.32913/rd-ict.vol1.no39.634
Biên tập lĩnh vực điều phối phản biện và quyết định nhận đăng: PGS. TS. Lê Hoàng Sơn

Tóm tắt: Phát hiện bộ phận cây từ ảnh là bước đầu tiên trong hệ thống nhận dạng cây. Các nghiên cứu gần đây thường dựa trên giả thuyết rằng loại bộ phận cây đã được xác định từ trước. Đã có một số nghiên cứu được đề xuất cho bài toán phát hiện tự động bộ phận cây nhưng các phương pháp này vẫn chủ yếu dựa trên các đặc trưng tự thiết kế. Trong bài báo này, chúng tôi đề xuất một phương pháp phát hiện tự động bộ phận cây sử dụng mạng nơ-ron tích chập. Các thực nghiệm được tiến hành trên tập con của tập dữ liệu PlantClef 2015 để đánh giá hiệu quả của phương pháp đề xuất. Phương pháp đề xuất cải thiện được 27,44% (đối với trường hợp bảy bộ phận) và 27,69% (đối với trường hợp năm bộ phận) tại hạng 1 so với phương pháp trước đó.

Từ khóa: *Phát hiện bộ phận cây, nhận dạng cây, học sâu, mạng nơ-ron tích chập.*

Title: **Automatic Plant Organ Detection from Images using Convolutional Neural Networks**

Abstract: Detecting plant organs from multiple organ images is the first step in a plant identification system. The current researches mainly rely on the assumption that the type of an organ is manually predetermined. Few works have been done on automatic plant organ detection but they are mainly based on hand-designed features. In this paper, we propose a method for automatic plant organ detection using the convolutional neural network. Different experiments on a subset of the PlantClef 2015 have been conducted to evaluate the robustness of the proposed method. The proposed method obtains 27.44% (for seven-organ cases) and 27.69% (for five-organ cases) of improvment in rank-1 over the state-of-the-art work.

Keywords: *Organ detection, plant identification, deep learning, convolutional neural networks.*

I. GIỚI THIỆU

Nhận dạng thực vật (loài cây) sử dụng ảnh của một hoặc nhiều bộ phận của cây đã và đang nhận được sự quan tâm của các nhà nghiên cứu trong các lĩnh vực phân loại thực vật học, đa dạng sinh học, tìm kiếm đa phương tiện, và thị giác máy tính. Ở khía cạnh của nhà nghiên cứu thực vật, công cụ tự động phân loại thực vật sử dụng ảnh các bộ phận cho phép cải thiện truy vấn trong nghiên cứu về đa dạng sinh học, cân bằng hệ sinh thái, khám phá dược phẩm, nhiên liệu, v.v. Đối với người dân, hàng ngày, mỗi người chúng ta tiếp xúc với rất nhiều cây, có nhiều cây gặp đi gặp lại nhiều lần, nhưng sự hiểu biết về cây còn hạn chế. Một công cụ tự động nhận biết cây trợ giúp cung cấp các thông tin như đặc điểm sinh họ và, công dụng là rất có ích. Trong nghiên cứu này, chúng tôi tập trung vào bài toán phân loại bộ phận cây từ hình ảnh. Việc phân loại bộ phận cây tự động này sẽ trợ giúp hoàn thiện hệ thống tự động nhận dạng cây dựa trên ảnh nhiều bộ phận có độ chính xác cao.

Một số hệ thống đã được triển khai và sử dụng rộng rãi như hệ thống Pl@ntnet [1], Leafsnap [2], MOSIR [3]. Trong các bộ phận của cây, lá là bộ phận thường được sử dụng để nhận dạng do bộ phận này dễ thu thập trong cả năm và thường có cấu trúc phẳng [4, 5]. Sau lá, hoa cũng được sử dụng để nhận dạng các loài do khả năng phân biệt cao. Tuy nhiên hoa chỉ nở theo mùa, tồn tại trong thời gian ngắn và có cấu trúc ba chiều [6]. Ngoài lá và hoa, các bộ phận khác như quả, thân hay toàn bộ cây cũng được sử dụng. Việc sử dụng một bộ phận thường không đầy đủ thông tin để nhận dạng một loài do sự tương tự lớn giữa các loài khác nhau và sự khác biệt giữa các ảnh cùng một bộ phận của cùng một loài.

Hình 1. Một số ảnh và tên bộ phận của cây trong PlantClef 2015 [9].

Hình 2. Các ảnh gây nhầm lẫn giữa các bộ phận. Chữ đậm bên dưới hình là tên của bộ phận được cung cấp bởi PlantClef 2015 [9].

Các nghiên cứu gần đây hướng tới việc nhận dạng cây dựa trên nhiều bộ phận của cây [7, 8]. Có bảy bộ phận được quan tâm: lá (ảnh lá trên nền phức tạp hoặc chụp trên cây), lá trên nền đơn giản (ảnh lá được tách khỏi cây và chụp trên nền đồng nhất), hoa, quả, thân, cành và toàn bộ cây. Các kết quả đã chỉ ra rằng việc kết hợp nhiều bộ phận cho phép nâng cao độ chính xác của các phương pháp nhận dạng cây dựa trên hình ảnh [7, 8]. Tuy nhiên, các nghiên cứu hiện tại thường dựa trên giả thuyết là kiểu bộ phận của cây đã được xác định từ trước - dựa trên việc gán nhãn thủ công. Đây là công việc rất tốn thời gian, đặc biệt là khi số lượng ảnh nhiều. Trong bài báo này chúng tôi đề xuất một phương pháp cho phép phát hiện tự động bộ phận của cây dựa trên ảnh chụp.

Nhận dạng tự động bộ phận gặp nhiều thách thức do các bộ phận có thể bị nhận nhầm lẫn nhau, đặc biệt với các ảnh được chụp trên nền phức tạp. Ngoài ra, trong một ảnh có thể có nhiều bộ phận khác nhau, gây nên sự khó khăn trong việc quyết định ảnh thuộc bộ phận nào. Hình 1 minh họa một số ảnh của bảy bộ phận trong PlantClef 2015 [9]. Hình 2 minh họa một số trường hợp khó do có nhiều bộ phận trên cùng một ảnh.

Để giải quyết cho những thách thức trên, hướng nghiên cứu trong bài báo là tìm cách thể hiện hiệu quả các đặc trưng của các bộ phận, trong đó các đặc trưng được học từ chính dữ liệu của ảnh bộ phận cây. Gần đây, các mạng nơ-ron tích chập (CNN: Convolutional neural network) [10, 11] đã chứng tỏ hiệu quả trong việc học các đặc trưng (trực quan) thông qua đáp ứng của các bộ lọc ở rất nhiều mức ngữ nghĩa khác nhau. Việc vận dụng các mạng CNN đã thành công ở các bài toán phân loại ảnh (Imagenet) [10], nhận dạng số/chữ viết [12]... Trong nghiên cứu này, các mạng CNN sẽ được làm thích nghi và các đặc trưng trích chọn từ mạng CNN sẽ được đánh giá cho bài toán nhận dạng bộ phận cây.

Đóng góp chính của bài báo là đề xuất một phương pháp phát hiện tự động các bộ phận dựa trên mạng nơ-ron tích chập. Phương pháp này được đánh giá thử nghiệm trên cơ sở dữ liệu PlantClef 2015 [9]. Các phương pháp dựa trên mạng nơ-ron tích chập thường thực hiện theo hai cách: (1) sử dụng đặc trưng và bộ phân lớp mặc định ở lớp kết nối đầy đủ; (2) trích chọn đặc trưng trước lớp cuối và đưa vào một bộ phân lớp. Trong nghiên cứu này, chúng tôi thực hiện đánh giá và so sánh hai cách tiếp cận trên nhằm xác định bộ phân lớp tốt nhất. Ngoài ra, chúng tôi cũng thực hiện đánh giá và hiển thị trực quan ba cấu hình mạng nổi tiếng (AlexNet, GoogLeNet và VGG [10, 13]), cũng như hai chiến lược khởi tạo trọng số (ngẫu nhiên, huấn luyện từ một cơ sở dữ liệu lớn hơn). Nhằm làm rõ hiệu quả của phương pháp đề xuất, chúng tôi thực hiện cài đặt so sánh kết quả phát hiện bộ phận của phương pháp đề xuất với phương pháp đã có trước đó dựa trên đặc trưng được thiết kế từ trước là đặc trưng GIST, và bộ phân lớp véc-tơ máy hỗ trợ (SVM: Support vector machine) [2, 14]. Mã nguồn của các phương pháp được cung cấp miễn phí cho cộng đồng nghiên cứu[1]. Cuối cùng, các kết quả nghiên cứu cho phép đưa ra gợi ý về số bộ phận cần sử dụng khi xây dựng cơ sở dữ liệu ảnh cho bài toán nhận dạng tự động cây.

II. NGHIÊN CỨU LIÊN QUAN

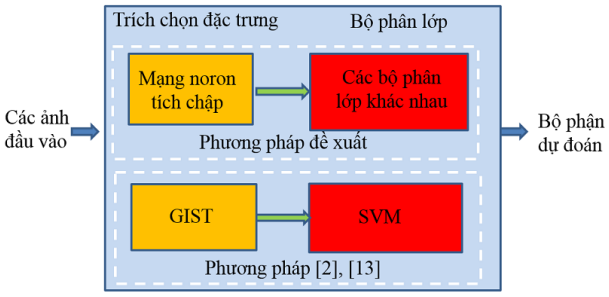
Hiện nay, các cơ sở dữ liệu cây thường dựa trên giả thuyết là các bộ phận của cây đã được xác định từ trước. Việc xác định bộ phận của cây thường thực hiện bằng phương pháp thủ công, nghĩa là người dùng chỉ ra loại bộ phận của cây có trong ảnh. Việc này đòi hỏi nhiều công sức và phụ thuộc vào chủ quan của người dùng. Cơ sở dữ liệu PlantClef từ năm 2015 [9] chứa dữ liệu ảnh các bộ phận của cây với thông tin bộ phận được xác định dựa trên việc gán nhãn thủ công bởi người dùng. Với mỗi ảnh, người dùng sẽ lựa chọn một trong bảy bộ phận. Hệ thống tra cứu cây Pl@ntnet [1] là ứng dụng đầu tiên nhận dạng cây dựa trên nhiều ảnh bộ phận. Tuy nhiên, khi người dùng đưa một ảnh truy vấn thì hệ thống yêu cầu chỉ rõ tên bộ phận có trong ảnh truy vấn [1].

Nhằm giảm thiểu yêu cầu đối với người dùng, một số nghiên cứu cho phép phát hiện tự động bộ phận dựa trên hình ảnh đã được đề xuất [2, 14, 15]. Trong [15], các tác

[1]http://www.mica.edu.vn/perso/Le-Thi-Lan/plant-organ-detection.html.

giả đề xuất sử dụng GIST như một bộ mô tả các đặc trưng và bộ phân lớp k láng giềng gần nhất (k-NN: k-nearest neighbors) để xác định bộ phận lá ở trong ảnh. Các tác giả thực hiện đánh giá trên cơ sở dữ liệu Flavia [16] gồm 32 lớp và đạt được độ chính xác 95%. Trong [2] và [14], các tác giả cũng đề xuất sử dụng bộ mô tả GIST, nhưng thay vì sử dụng thuật toán k-NN, SVM được đề xuất sử dụng để xác định bộ phận lá ở trong ảnh. Phương pháp trình bày trong [2] được đánh giá trên cơ sở dữ liệu Leafsnap [17] gồm 5.972 ảnh với độ chính xác 62,9%. Kết quả phân lớp trong [14] đạt độ chính xác 98,67% trên cơ sở dữ liệu được xây dựng từ hệ thống Leafsnap kết hợp với công cụ tìm kiếm Google và tự thu thập. Các phương pháp đề xuất trong [2, 14] đạt được kết quả cao trên cơ dữ liệu thử nghiệm. Tuy nhiên, các phương pháp [2, 14] chỉ nhằm xác định một bộ phận duy nhất (lá cây) có trong ảnh hay không (phân lớp nhị phân) mà chưa quan tâm đến phân lớp nhiều bộ phận khác nhau (phân lớp nhiều lớp). Ngoài ra, các ảnh trong các cơ sở dữ liệu thử nghiệm trong [2, 14] là các ảnh lá cây chụp trên nền đơn giản. Theo hiểu biết của chúng tôi, chưa có nghiên cứu nào về bài toán xác định nhiều bộ phận của cây và thực hiện trên các cơ sở dữ liệu đa dạng và phức tạp.

Trong những năm gần đây, phương pháp học sâu phát triển rất nhanh dựa trên lượng dữ liệu huấn luyện lớn và khả năng tính toán ngày càng mạnh của các máy tính. Trong lĩnh vực thị giác máy tính, mạng nơ-ron tích chập với khả năng tự học các đặc trưng đã chứng minh hiệu quả trong các bài toán phát hiện và phân loại đối tượng [18], với một số mạng nổi tiếng như AlexNet, VGG, GoogLeNet. Các mạng nơ-ron tích chập này cũng đã được áp dụng cho bài toán nhận dạng cây, đặc biệt trong cuộc thi PlantClef từ năm 2014 đến năm 2017, và cho các kết quả rất tốt so với các phương pháp truyền thống sử dụng các đặc trưng được thiết kế từ trước [7, 19, 20]. Tuy nhiên theo hiểu biết của chúng tôi, chưa có một nghiên cứu nào áp dụng mạng nơ-ron tích chập cho bài toán phát hiện bộ phận cây cũng như so sánh đánh giá giữa cách tiếp cận truyền thống (dựa trên trích chọn đặc trưng thiết kế) và phương pháp dựa trên cách tiếp cận học sâu. Do vậy trong bài báo này, chúng tôi triển khai phương pháp phát hiện tự động các bộ phận của cây dựa trên mạng nơ-ron tích chập. Các kết quả thử nghiệm trên cơ sở dữ liệu gồm 235 loài từ PlantClef 2015 được so sánh với [2, 14] chứng tỏ hiệu quả của phương pháp đề xuất. Ngoài ra, các phương pháp gần đây tập trung nâng cao độ chính xác nhận dạng sử dụng ảnh của cây thường mặc định sử dụng nhãn các bộ phận được phân loại trước (thủ công) [1, 7, 20]. Cách tiếp cận trong bài báo mở ra hướng giải quyết cho bài toán phân loại tự động hoàn toàn từ quá trình xác định các bộ phận, đến quá trình nhận dạng cuối cùng.

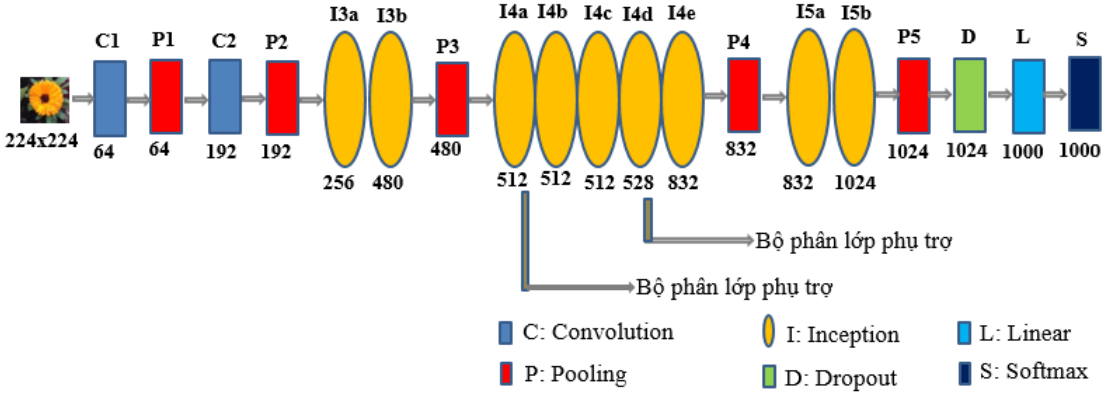


Hình 3. Phương pháp đề xuất và sự so sánh tương ứng với phương pháp [2, 14].

III. PHƯƠNG PHÁP ĐỀ XUẤT

Bài toán xác định tự động các bộ phận là bài toán xác định một ảnh x thuộc vào một trong C bộ phận. C gồm bảy bộ phận như trong định nghĩa của PlantClef. Hình 3 minh họa phương pháp đề xuất và so sánh tương ứng giữa phương pháp đề xuất và phương pháp [2, 14]. Từ ảnh đầu vào, áp dụng mạng nơ-ron tích chập đề xuất để trích ra véc tơ đặc trưng, sau đó đưa vào các bộ phân lớp khác nhau.

Với bài toán phân loại các bộ phận của cây, hiện chưa có nghiên cứu nào đề xuất sử dụng mạng nơ-ron tích chập, bởi vậy trong bài báo này chúng tôi đề xuất sử dụng mạng GoogLeNet do các kết quả ấn tượng của mạng này cho các bài toán phân lớp đối tượng. Mạng GoogLeNet do Szegedy và các cộng sự đề xuất [13] đã đứng vị trí thứ nhất của cuộc thi nhận dạng hình ảnh quy mô lớn (ILSVRC) trong năm 2014, từ đó đến nay mạng này đã được sử dụng nhiều cho các bài toán phát hiện và nhận dạng. GoogLeNet là kiến trúc đầu tiên giới thiệu mô đun inception, cho phép làm giảm một số lượng lớn các tham số huấn luyện trong mạng. Mô đun inception sử dụng việc kết hợp song song các lớp nhân chập (Convolution) 1×1, 3×3, 5×5 với các lớp giảm chiều (Pooling). Kiến trúc này còn được gọi là mạng trong mạng. Kiến trúc GoogLeNet là mạng sâu với 22 lớp khi chỉ tính các lớp có chứa tham số, lớp trên cùng là hàm phân lớp Softmax. Mạng GoogLeNet sử dụng kiến trúc sâu hơn và rộng hơn so với nhiều mạng nơ-ron tích chập khác như AlexNet, VGG. Kiến trúc thông thường của một mạng nơ-ron tích chập thường bao gồm nhiều lớp theo cấu trúc (một vài lớp nhân chập theo sau là lớp giảm chiều) sau cùng là các lớp kết nối đầy đủ. Hình 4 chỉ ra kiến trúc của mạng GoogLeNet với chín mô đun inception và số đầu ra tương ứng của mỗi lớp. Trong đó, ký hiệu C_i, P_i, I_{ij}, D, L và S được sử dụng với ý nghĩa như sau: C, P, I, D, L và S là viết tắt tương ứng của lớp nhân chập, lớp giảm chiều, mô đun Inception, lớp Dropout, lớp Linear, lớp Softmax, $i = \{1,2,3,4,5\}$ là chỉ mục của lớp đang xét, $j = \{a,b,c,d,e\}$ là chỉ mục của các mô đun Inception khác nhau trong cùng một lớp. Đầu ra của các lớp nhân chập và



Hình 4. Kiến trúc của mạng GoogLeNet [13] với số đầu ra của mỗi lớp được thể hiện dưới mỗi lớp

lớp giảm chiều là các véc tơ đặc trưng. Các đặc trưng thu được ở các lớp sau thì càng trừu tượng hơn các đặc trưng thu được ở lớp trước. Trong bước này chúng tôi sẽ trích rút đặc trưng sau lớp P5, là lớp giảm chiều trung bình có tên là pool5/7× 7_s1, ở đầu ra tại lớp này thu được véc tơ 1024 chiều. Lớp này trích rút được các đặc trưng mức cao nhất của ảnh và cung cấp các thông tin mô tả tốt nhất về các đối tượng trong ảnh. Cách tính số chiều véc tơ đặc trưng thu được ở lớp này như sau: cho ảnh đầu vào có kích thước 224×224, khi đi qua các lớp, do phụ thuộc vào số bộ lọc, kích thước bộ lọc, các tham số dịch chuyển bộ lọc của mỗi lớp, sẽ thu được các véc tơ đặc trưng đầu ra có số chiều như sau:

$$
\begin{aligned}
&Input\_image \rightarrow C1 \rightarrow 112 \times 112 \times 64 \rightarrow P1 \rightarrow \\
&56 \times 56 \times 64 \rightarrow C2 \rightarrow 56 \times 56 \times 192 \rightarrow P2 \rightarrow \\
&28 \times 28 \times 192 \rightarrow I3a \rightarrow 28 \times 28 \times 256 \rightarrow I3b \rightarrow \\
&28 \times 28 \times 480 \rightarrow P3 \rightarrow 14 \times 14 \times 480 \rightarrow I4a \rightarrow \\
&14 \times 14 \times 512 \rightarrow I4b \rightarrow 14 \times 14 \times 512 \rightarrow I4c \rightarrow \\
&14 \times 14 \times 512 \rightarrow I4d \rightarrow 14 \times 14 \times 528 \rightarrow I4e \rightarrow \\
&14 \times 14 \times 832 \rightarrow P4 \rightarrow 7 \times 7 \times 832 \rightarrow I5a \rightarrow \\
&7 \times 7 \times 832 \rightarrow I5b \rightarrow 7 \times 7 \times 1024 \rightarrow P5 \\
&\rightarrow 1 \times 1 \times 1024
\end{aligned}
$$

Mặc dù việc sử dụng các mạng CNN ngày càng phổ biến và đạt hiệu quả cao trong các bài toán phân loại ảnh, hạn chế của việc sử dụng mạng CNN đối với một vấn đề nhận dạng mới là: (1) cơ sở dữ liệu huấn luyện thường phải lớn để học các đặc trưng ở nhiều lớp (layer) của mạng; (2) Việc huấn luyện mô hình mất nhiều thời gian. Để giải quyết vấn đề này, kỹ thuật học chuyển giao (transfer learning) sẽ được vận dụng. Theo kỹ thuật này, một mạng CNN đã được huấn luyện từ trước để giải quyết bài toán phân lớp trên bộ cơ sở dữ liệu đủ lớn và đa dạng. Trong nghiên cứu này, chúng tôi sử dụng mạng GoogLeNet đã được huấn luyện trên bộ cơ sở dữ liệu Imagenet chứa 1,2 triệu ảnh với 1000 lớp [21]. Bộ tham số của mạng này đã được tích hợp trong bộ công cụ thư viện sử dụng (Caffe Library [22]). Cần chú ý là mạng này không được sử dụng trực tiếp với bài toán phân lớp bảy bộ phận cây trong nghiên cứu. Thay vì đó, bộ tham số sẽ được sử dụng để khởi tạo mạng; sau đó sẽ được tinh chỉnh trên bộ cơ sở dữ liệu làm việc. Để thấy rõ vai trò của việc khởi tạo trọng số, chúng tôi thực hiện thêm thử nghiệm và so sánh độ chính xác trên cùng một cấu hình mạng với việc khởi tạo trọng số ngẫu nhiên và trọng số khởi tạo dựa trên cơ sở dữ liệu ImageNet.

Nhằm tăng sự đa dạng của dữ liệu, chúng tôi thực hiện mở rộng dữ liệu trong quá trình huấn luyện bằng phép lấy gương, điều chỉnh kích thước của ảnh về 240×240, sau đó xén ngẫu nhiên để đưa về kích thước 224×224. Việc mở rộng dữ liệu được áp dụng để làm giảm cơ hội học quá khớp trong quá trình huấn luyện và cải thiện kết quả phân loại trong quá trình kiểm thử. Để làm rõ ưu điểm của kiến trúc mạng GoogLeNet, chúng tôi đã thực hiện thêm thực nghiệm so sánh GoogLeNet với hai mạng điển hình khác là AlexNet và VGG-16.

IV. KẾT QUẢ THỰC NGHIỆM

Chúng tôi thực hiện thực nghiệm trên cơ sở dữ liệu PlantClef 2015 [9]. Dữ liệu này chứa 1000 loài, mỗi ảnh sẽ thuộc về một trong bảy bộ phận: lá, lá trên nền đơn giản, hoa, quả, cành, thân, toàn bộ cây. Tuy nhiên không phải loài nào cũng có các ảnh của đầy đủ cả bảy bộ phận trên. Vì vậy để phục vụ việc phân loại các bộ phận, chúng tôi đã lọc ra từ cơ sở dữ liệu này những loài có đầy đủ cả bảy bộ phận, kết quả thu được 235 loài (Bảng I).

Chúng tôi cài đặt GoogLeNet sử dụng Caffe [22], một nền tảng cho các phương pháp học sâu, với các trọng số tiền huấn luyện thu được từ Caffe Model Zoo học được từ cơ sở dữ liệu Imagenet. Các thực nghiệm được tiến hành trên máy chủ được trang bị 11 GB GPU.

Bảng I
THÔNG TIN CƠ SỞ DỮ LIỆU THỰC NGHIỆM

Tên bộ phận	**Tập huấn luyện**	**Tập kiểm thử**
Lá (Leaf)	7.666	1.589
Lá trên nền đơn giản (Leafscan)	11.365	209
Hoa (Flower)	8.035	1.970
Quả (Fruit)	4.022	835
Thân (Stem)	3.693	434
Cành (Branch)	3.643	955
Ảnh toàn bộ cây (Entire)	3.493	1.280
Tổng	41.917	7.272

Để đánh giá các kết quả thực nghiệm chúng tôi sử dụng độ đo độ chính xác $\text{Acc}_{\text{rank-}k}$ tại hạng thứ k, được định nghĩa như sau:

$$\text{Acc}_{\text{rank-}k} = \frac{T_{\text{rank-}k}}{N}, \quad (1)$$

trong đó $T_{\text{rank-}k}$ là số kết quả phát hiện đúng ở k vị trí đầu tiên trong kết quả trả về, N là tổng số các ảnh truy vấn. Các nghiên cứu trước đó thường đánh giá độ chính xác ở hạng 1 ($k = 1$). Trong nghiên cứu này, chúng tôi thấy rằng, với các ảnh phức tạp, thay vì việc đưa ra một bộ phận duy nhất, hệ thống có thể xem xét để đưa ra hai bộ phận tồn tại trong ảnh. Do đó, chúng tôi thực hiện đánh giá hệ thống ở cả hai hạng: hạng 1 ($k = 1$) và hạng 2 ($k = 2$). Chúng tôi đã thực hiện bốn thực nghiệm và đạt được các kết quả như trình bày dưới đây.

1. Thực nghiệm 1

Thực hiện phân loại bảy bộ phận theo mạng Googlenet. Các tham số được sử dụng như sau: kích thước bó = 32; tốc độ học = 0,0001. Trong thực nghiệm 1, chúng tôi sử dụng bộ phân lớp mặc định trong mạng nơ-ron tích chập (bộ phân lớp Softmax). Kết quả đạt độ chính xác tại hạng 1 và hạng 2 lần lượt là **82,60%** và **93,45%**. Kết quả nhận dạng này là khá cao khi số phân lớp ở đây là bảy, trong đó có sáu bộ phận chủ yếu có nền phức tạp. Điều này chứng tỏ kỹ thuật học sâu có khả năng học tốt với các ảnh tự nhiên.

Bảng II trình bày kết quả tương ứng với hai chiến lược khởi tạo trọng số: ngẫu nhiên và sử dụng bộ trọng số đã tiền huấn luyện trên ImageNet. Kết quả cho thấy, khi sử dụng bộ trọng số đã huấn luyện trên một cơ sở dữ liệu lớn hơn là ImageNet, độ chính xác tăng thêm 6,65% ở hạng 1 và 4,77% ở hạng 2.

Bảng III thể hiện ma trận nhầm lẫn (confusion matrix) tính theo phần trăm. Các bộ phận cho hiệu quả phát hiện từ cao xuống thấp là thân (92,4%), hoa (91,62%), lá trên nền đơn giản (89,0%), lá (87,35%), ảnh toàn bộ (84,3%),

Bảng II
SO SÁNH VIỆC ÁP DỤNG MẠNG GOOGLENET DỰA TRÊN BỘ TRỌNG SỐ KHỞI TẠO NGẪU NHIÊN VÀ BỘ TRỌNG SỐ TIỀN HUẤN LUYỆN TRÊN IMAGENET

Chiến lược khởi tạo trọng số	$\text{Acc}_{\text{rank-1}}$	$\text{Acc}_{\text{rank-2}}$
Khởi tạo ngẫu nhiên	74,05%	88,68%
Tiền huấn luyện trên ImageNet	**82,60%**	**93,45%**

quả (74,97%), cành (54,66%). Hình 5 minh họa một số ví dụ về các trường hợp nhận dạng nhầm giữa các bộ phận khác nhau. Từ việc phân tích các kết quả thu được, cho thấy một số lá trên nền đơn giản có thể bị nhận nhầm sang lá trong một số trường hợp nền không phải là màu trắng. Lá bị nhận nhầm thành thân trong một số trường hợp khi chụp ảnh lá với cự ly quá gần, do hệ thống nhận nhầm gân lá với thân. Ảnh hoa bị nhận nhầm sang lá trong trường hợp nụ hoa thon dài, ảnh có chứa lá đài của hoa, ảnh chụp ở cự ly xa, hoa nhỏ trong khi ảnh lá lại to; hoa bị nhận nhầm sang quả khi nụ hoa có hình dạng rất giống quả. Ảnh quả bị nhận nhầm sang hoa thường là với các ảnh quả dạng chùm và đối xứng giống hoa. Ảnh thân bị nhận nhầm sang một số bộ phận khác như lá, hoa và quả thường là những ảnh chụp có thân nhỏ, màu xanh, gắn kèm trên đó lá, hoa hay quả. Thân là bộ phận có khả năng phân biệt cao nhất do ảnh thân có các đặc trưng kết cấu, màu sắc rất dễ phân biệt với các bộ phận khác và ảnh chụp thường là không chứa bộ phận khác, đối tượng thân thường chiếm hết không gian ảnh. Cành có kết quả phân loại thấp nhất, là bộ phận dễ gây nhập nhằng nhất đối với các ảnh bộ phận khác vì ảnh cành thường có chứa cả lá, hoa, quả và thân.

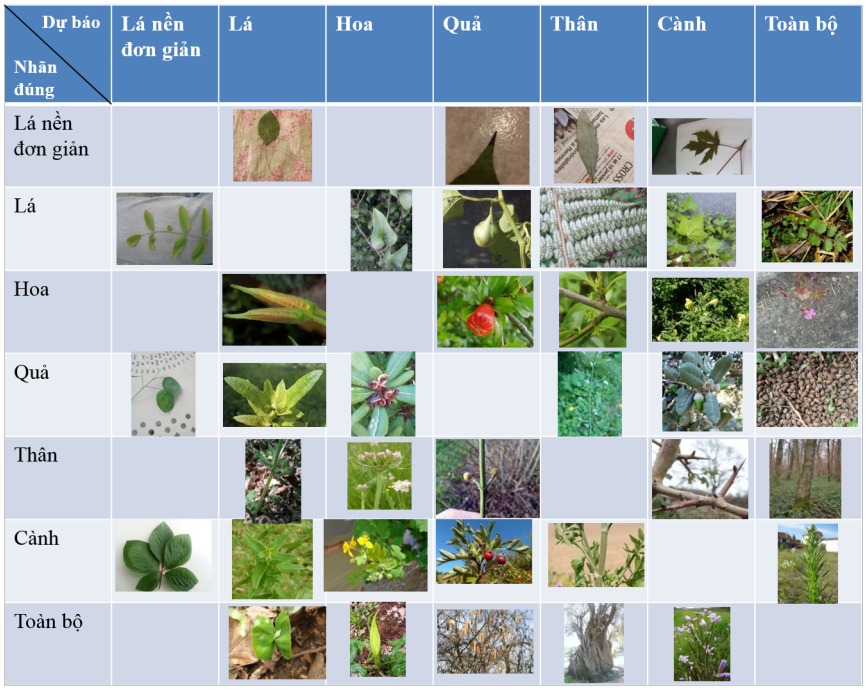
Kết quả nhận nhầm cũng xuất phát từ sự nhập nhằng và độ chính xác trong kết quả xác định bộ phận thủ công được cung cấp trong PlantClef2015.

Để làm rõ hiệu quả của cấu hình mạng lựa chọn, chúng tôi đã thực hiện so sánh kết quả phát hiện bộ phận với ba cấu hình mạng AlexNet, VGG-16 và GoogLeNet. Các độ chính xác ở hạng 1 là 81,19% cho AlexNet, 77,19% cho VGG-16, và 82,6% cho GoogLeNet. Mạng GoogLeNet cho kết quả tốt nhất do mạng này có kiến trúc sâu hơn, rộng hơn các mạng AlexNet và VGG-16.

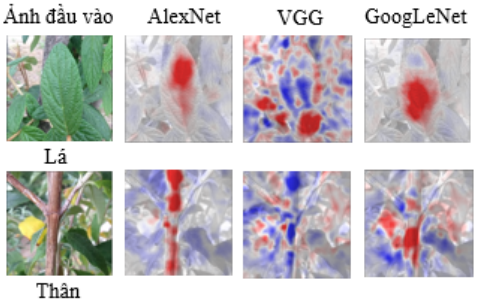
Ngoài ra, để hiển thị trực quan quyết định nhận dạng của các mạng, chúng tôi áp dụng phương pháp biểu diễn trong bài báo [23]. Hình 6 chỉ ra các kết quả của 3 mạng khác trên 2 ảnh đầu vào, vùng màu đỏ thể hiện vùng dự đoán tin cậy, trong khi vùng màu xanh thể hiện vùng dự đoán không tin cậy. Kết quả cho thấy AlexNet và GoogLeNet thể hiện rất rõ các vùng dự đoán ở phần trung tâm của đối tượng, trong khi VGG lại không tập trung vào trung tâm của đối tượng mà rải rác ở nhiều phần quanh đối tượng, và quan tâm đến vùng nền của đối tượng.

Bảng III
MA TRẬN NHẦM LẪN PHÁT HIỆN BẢY BỘ PHẬN

		Dự báo						
		Lá nền đơn giản	Lá	Hoa	Quả	Thân	Cành	Toàn bộ
Nhãn đúng	Lá nền đơn giản	**89,0**	8,61	0,0	0,96	0,48	0,96	0,0
	Lá	0,88	**87,35**	1,01	1,57	0,44	7,43	1,32
	Hoa	0,0	0,36	**91,62**	2,34	0,1	3,65	1,93
	Quả	0,36	1,68	10,54	**74,97**	0,6	10,3	1,56
	Thân	0,0	0,46	0,46	1,15	**92,4**	2,53	3,0
	Cành	0,73	10,79	11,52	5,97	0,73	**54,66**	15,6
	Toàn bộ	0,0	3,2	2,73	0,78	0,39	8,59	**84,3**



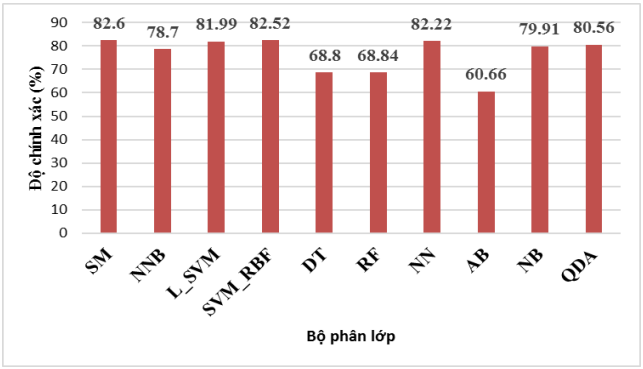
Hình 5. Một số ví dụ minh họa về các trường hợp nhận dạng nhầm giữa các bộ phận khác nhau.



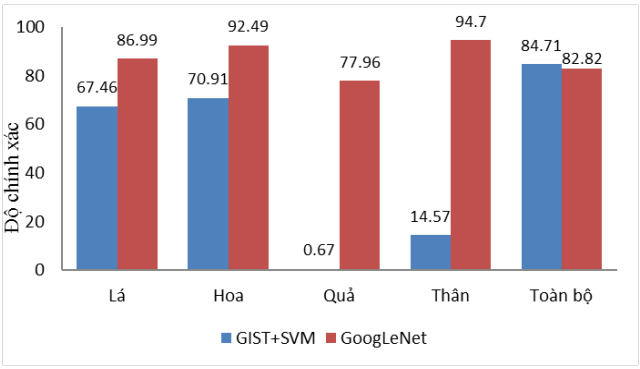
Hình 6. So sánh trực quan các dự báo của các kiến trúc mạng khác nhau: AlexNet, GoogLeNet và VGG-16. Vùng tin cậy cho dự đoán được hiển thị bằng màu đỏ, vùng dự đoán không tin cậy có màu xanh.

2. Thực nghiệm 2

Với mục đích đánh giá các bộ phân lớp khác nhau trên cùng bộ đặc trưng được trích rút từ mạng nơ-ron tích chập, chúng tôi trích rút lớp đặc trưng cuối cùng trước lớp kết nối đầy đủ và cho qua các bộ phân lớp khác nhau: láng giềng gần nhất (NNB: Nearest neighbors), máy véc tơ hỗ trợ tuyến tính (L_SVM: Linear SVM), máy véc tơ hỗ trợ phi tuyến sử dụng nhân RBF (SVM_RBF), cây quyết định (DT: Decision tree), rừng ngẫu nhiên (RF: Random forest), mạng nơ-ron (NN: Neural network), Bayes thô (NB: Naïve Bayes), phân tích khác biệt cầu phương (QDA: Quadratic discriminant analysis) để so sánh với bộ phân lớp Softmax (SM) của mạng GoogLeNet. Các bộ phân lớp này được xét cho bảy bộ phận.



Hình 7. Độ chính xác phát hiện các bộ phận ở hạng 1 với các bộ phân lớp khác nhau.



Hình 8. So sánh kết quả của phương pháp đề xuất và phương pháp trong [2, 14] trên năm bộ phận.

Hình 7 chỉ ra kết quả trên các bộ phân lớp khác nhau, các kết quả được xếp lần lượt từ cao xuống thấp như sau: SM (82,6%), SVM_RBF (82,52%), NN (82,22%), L_SVM (81,99%), QDA (80,56%), NB (79,91%), NNB (78,7%), RF (68,84%), DT (68,8%), AB (60,66%). Bộ phân lớp SM của chính mạng GoogLeNet cho kết quả tốt nhất là 82,6% đối với bài toán phân loại bảy bộ phận của cây, các đặc trưng này là phù hợp với bộ phân lớp Softmax. Bộ phân lớp máy véc tơ hỗ trợ phi tuyến sử dụng nhân RBF, mạng nơ-ron và bộ phân lớp máy véc tơ hỗ trợ tuyến tính cho các kết quả khá gần với bộ phân lớp Softmax.

3. Thực nghiệm 3

Các kết quả phân tích ở thực nghiệm 1 cho thấy, việc phân chia thành bảy bộ phận là không hợp lý do có sự tương tự và nhập nhằng trong việc xác định các bộ phận có trong một ảnh. Chúng tôi đề xuất một tập gồm năm bộ phận thay vì bảy bộ phận bằng cách nhóm các bộ phận tương tự nhau. Năm bộ phận được quan tâm là: lá (bao gồm lá chụp trên các loại nền khác nhau), hoa, quả, thân và toàn bộ (bao gồm ảnh toàn bộ cây và cành cây). Chúng tôi đánh giá phương pháp đề xuất trên năm bộ phận này. Độ chính xác thu được ở hạng 1 và hạng 2 lần lượt là 86,62% và 97,08%.

4. Thực nghiệm 4

Để so sánh giữa phương pháp học sâu với phương pháp đề xuất trong [2, 14], chúng tôi cài đặt và thử nghiệm lại các phương pháp này trên cùng cơ sở dữ liệu thử nghiệm. Từ một ảnh đầu vào, đặc trưng GIST gồm 512 chiều sẽ được trích rút. Sau đó, bộ phân lớp máy véc tơ hỗ trợ (SVM) được áp dụng. Kết quả đạt được độ chính xác 55,16%, thấp hơn 27,44% so với việc áp dụng mạng Googlenet với bộ phân lớp Softmax, và thấp hơn tất cả các bộ phân lớp khác ở thực nghiệm 2. Điều này cho thấy phương pháp học sâu hiệu quả hơn nhiều so với cách tiếp cận truyền thống cho bài toán phát hiện các bộ của phận cây, đặc biệt là khi ảnh thu được trong các điều kiện phức tạp.

Chúng tôi cũng áp dụng cách làm này đối với năm bộ phận như ở thực nghiệm 3, kết quả đạt độ chính xác là 58,93%, thấp hơn so với phương pháp đề xuất (86,62%). Hình 8 thể hiện so sánh kết quả của phương pháp đề xuất và phương pháp [2, 14] cho từng bộ phận. Các bộ phận lá, hoa, thân và quả sử dụng mạng GoogLeNet cho kết quả cao hơn hẳn với phương pháp sử dụng GIST và SVM [2, 14]. Phương pháp [2, 14] đạt độ chính xác 0,67% cho bộ phận quả do quả chiếm một ví trí nhỏ trong ảnh trong khi đặc trưng GIST là đặc trưng toàn cục. Một điểm thú vị là đối với ảnh toàn bộ cây thì phương pháp [2, 14] cho kết quả cao hơn phương pháp đề xuất 1,81% do ảnh toàn bộ cây thường chiếm không gian toàn bộ ảnh, màu sắc trong ảnh chủ yếu là màu xanh. Đặc trưng GIST có khả năng trích chọn đặc điểm đó và phân biệt ảnh toàn bộ cây.

V. KẾT LUẬN

Bài báo này đã đề xuất sử dụng mạng nơ-ron tích chập GoogLeNet cho việc phát hiện các bộ phận của cây với độ chính xác theo hạng 1 và hạng 2 lần lượt là 82,6%, 93,45% đối với trường hợp bảy bộ phận, và lần lượt là 86,62% và 97,08% đối với trường hợp năm bộ phận. Các kết quả cho thấy phương pháp đề xuất cải thiện độ chính xác ở hạng 1 so với phương pháp ở [2, 14] là 27,44% cho bảy bộ phận và 27,69% cho năm bộ phận. Các kết quả trong các thực nghiệm cũng cho thấy vai trò của việc khởi tạo trọng số của các mạng, cũng như hiệu quả của mạng GoogLeNet so với mạng VGG-16 và AlexNet cho bài toán nhận dạng các bộ phận. Ngoài ra, các kết quả hiển thị cho phép giải thích tường minh các kết luận nhận dạng của các mạng. Các kết quả thử nghiệm trong bài báo giúp đưa ra gợi ý về việc lựa chọn số bộ phận của cây trong quá trình xây dựng cơ sở dữ liệu hình ảnh phục vụ cho bài toán nhận dạng tự động cây từ hình ảnh.

Trong tương lai chúng tôi sẽ tiếp tục nghiên cứu để cải tiến kết quả phát hiện tự động các bộ phận theo hướng kết hợp cả mạng nơ-ron và các đặc trưng thiết kế trước, đồng thời thực hiện dự báo nhãn của loài dựa trên bộ phận cây đã phát hiện được.

TÀI LIỆU THAM KHẢO

[1] A. Joly, H. Goëau, P. Bonnet, V. Bakić, J. Barbe, S. Selmi, I. Yahiaoui, J. Carré, E. Mouysset, J.-F. Molino *et al.*, "Interactive plant identification based on social image data," *Ecological Informatics*, vol. 23, pp. 22–34, 2014.

[2] N. Kumar, P. N. Belhumeur, A. Biswas, D. W. Jacobs, W. J. Kress, I. C. Lopez, and J. V. Soares, "Leafsnap: A computer vision system for automatic plant species identification," in *Computer Vision–ECCV 2012*. Springer, 2012, pp. 502–516.

[3] K. H. Phyu, A. Kutics, and A. Nakagawa, "Self-adaptive feature extraction scheme for mobile image retrieval of flowers," in *Proceedings of the Eighth International Conference on Signal Image Technology and Internet Based Systems (SITIS)*. IEEE, 2012, pp. 366–373.

[4] J. S. Cope, D. Corney, J. Y. Clark, P. Remagnino, and P. Wilkin, "Plant species identification using digital morphometrics: A review," *Expert Systems with Applications*, vol. 39, no. 8, pp. 7562–7573, 2012.

[5] P. Bonnet, A. Joly, H. Goëau, J. Champ, C. Vignau, J.-F. Molino, D. Barthélémy, and N. Boujemaa, "Plant identification: man vs. machine," *Multimedia Tools and Applications*, vol. 75, no. 3, pp. 1647–1665, 2016.

[6] M.-E. Nilsback and A. Zisserman, "Automated flower classification over a large number of classes," in *Proceedings of the Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP'08)*. IEEE, 2008, pp. 722–729.

[7] H. Goëau, P. Bonnet, and A. Joly, "LifeCLEF Plant Identification Task 2015," in *Proceedings of the Conference and Labs of the Evaluation forum (CLEF)*, ser. CLEF2015 Working notes, CEUR-WS, Ed., vol. 1391, Toulouse, France, Sep. 2015. [Online]. Available: https://hal.inria.fr/hal-01182795

[8] T. T.-N. Nguyen, T.-L. Le, H. Vu, H.-H. Nguyen, and V.-S. Hoang, "A combination of deep learning and hand-designed feature for plant identification based on leaf and flower images," in *Advanced Topics in Intelligent Information and Database Systems*. Springer, 2017, pp. 223–233.

[9] "http://www.imageclef.org/lifeclef/2015/plant, (retrieved 30/8/2015)."

[10] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Proceedings of the Advances in neural information processing systems*, 2012, pp. 1097–1105.

[11] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2014, pp. 580–587.

[12] Y. LeCun, L. Bottou, Y. Bengio, and P. Haffner, "Gradient-based learning applied to document recognition," *Proceedings of the IEEE*, vol. 86, no. 11, pp. 2278–2324, 1998.

[13] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, 2015, pp. 1–9.

[14] Phan Thị Thu Hồng, Đoàn Thị Thu Hà, and Nguyễn Thị Thủy, "Ứng dụng phân lớp ảnh chụp lá cây bằng phương pháp máy vecto hỗ trợ," *Tạp chí khoa học và phát triển*, vol. 11, no. 7, pp. 1045–1052, 2013.

[15] Q.-K. Nguyen, T.-L. Le, and N.-H. Pham, "Leaf based plant identification system for android using surf features in combination with bag of words model and supervised learning," in *Proceedings of the International Conference on Advanced Technologies for Communications (ATC)*. IEEE, 2013, pp. 404–407.

[16] "http://flavia.sourceforge.net/, (retrieved 10/9/2018)."

[17] "http://leafsnap.com/dataset/, (retrieved 15/10/2018)."

[18] H.-J. Yoo, "Deep convolution neural networks in computer vision," *IEIE Transactions on Smart Processing & Computing*, vol. 4, no. 1, pp. 35–43, 2015.

[19] H. Goëau, P. Bonnet, and A. Joly, "Plant identification in an open-world (lifeclef 2016)," *CLEF working notes*, vol. 2016, pp. 428–439, 2016.

[20] H. Goeau, P. Bonnet, and A. Joly, "Plant identification based on noisy web data: the amazing performance of deep learning (lifeclef 2017)." CEUR Workshop Proceedings, 2017.

[21] "http://www.image-net.org/download-images, (retrieved 5/11/2018)."

[22] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell, "Caffe: Convolutional architecture for fast feature embedding," in *Proceedings of the 22nd ACM international conference on Multimedia*. ACM, 2014, pp. 675–678.

[23] L. M. Zintgraf, T. S. Cohen, T. Adel, and M. Welling, "Visualizing deep neural network decisions: Prediction difference analysis," *CoRR*, vol. abs/1702.04595, 2017.

Nguyễn Thị Thanh Nhàn sinh năm 1981 tại Bắc Giang. Tác giả tốt nghiệp Trường Đại học Khoa học Tự nhiên, Đại học Quốc gia Hà Nội năm 2003 và nhận bằng Thạc sĩ năm 2007, tại Đại học Thái Nguyên. Hiện nay, tác giả là giảng viên tại Khoa Công nghệ thông tin, Trường Đại học Công nghệ thông tin và Truyền thông Thái Nguyên và là nghiên cứu sinh tại Trường Đại học Bách Khoa Hà Nội. Lĩnh vực nghiên cứu của tác giả là xử lý ảnh, thị giác máy, nhận dạng.

Lê Thị Lan nhận bằng Tiến sĩ chuyên ngành Xử lý ảnh tại Đại học Nice, Cộng hòa Pháp, năm 2009. Hiện nay, tác giả là giảng viên phòng Thị giác máy tính, Viện nghiên cứu quốc tế MICA, Trường Đại học Bách khoa Hà Nội. Các lĩnh vực nghiên cứu của tác giả là tìm kiếm thông tin ảnh và video dựa trên nội dung, phân tích và hiểu nội dung ảnh và video, tương tác người - máy.

Vũ Hải nhận bằng Tiến sĩ chuyên ngành Khoa học máy tính tại Trường Đại học Osaka, Nhật Bản, năm 2009. Hiện nay, ông là giảng viên tại phòng Thị giác máy tính, Viện Nghiên cứu quốc tế MICA, Trường Đại học Bách khoa Hà Nội. Các lĩnh vực nghiên cứu quan tâm của ông bao gồm phân tích ảnh y tế hỗ trợ chuẩn đoán, đặc biệt ảnh nội soi không dây; thị giác máy tính trong robotics và trong nông nghiệp.

Hoàng Văn Sâm nhận bằng Tiến sĩ chuyên ngành Phân loại thực vật và bảo tồn Đa dạng sinh học tại Đại học Leiden, Hà Lan, năm 2009. Ông được phong Phó giáo sư ngành Lâm nghiệp năm 2013. Hiện nay, ông là giảng viên cao cấp Bộ môn Thực vật rừng, Trường Đại học Lâm nghiệp. Lĩnh vực nghiên cứu của ông bao gồm phân loại thực vật, bảo tồn đa dạng sinh học, quản lý vườn quốc gia, khu bảo tồn thiên nhiên.