HIỂU BIẾT TOÀN DIỆN VỀ BIỂU DIỄN ĐẶC TRƯNG CỤC BỘ (SIFT) VÀ TOÀN CỤC (VGG) CHO PHÂN LOẠI HÌNH ẢNH

Huỳnh Văn Nguyên Bảo, Lê Quang Chiến

Khoa Công nghệ Thông tin, Trường Đại học Khoa học, Đại học Huế

Email: hvnguyenbao0611@gmail.com, lqchien@husc.edu.vn

Ngày nhận bài: 12/6/2024; ngày hoàn thành phản biện: 24/6/2024; ngày duyệt đăng: 01/11/2024

TÓM TẮT

Phân loại hình ảnh là một trong những bài toán quan trọng trong lĩnh vực thị giác máy tính. Bài báo này đánh giá hai phương pháp chính trong phân loại hình ảnh: đặc trưng cục bộ và đặc trưng toàn cục. Kỹ thuật cổ điển như SIFT tập trung vào đặc trưng cục bộ, trong khi phương pháp hiện đại như Convolutional Neural Network (CNN), tiêu biểu là VGG-8, khai thác đặc trưng toàn cục. Các thí nghiệm được thực hiện nhằm so sánh hiệu suất của hai phương pháp trên các tập dữ liệu thực tế. Kết quả thu được cung cấp một đánh giá toàn diện về khả năng phân loại hình ảnh và ứng dụng của từng phương pháp. Điều này giúp xác định ưu và nhược điểm của mỗi phương pháp trong các bối cảnh khác nhau.

Từ khóa: Thị giác máy tính, phân loại hình ảnh, đặc trưng cục bộ, đặc trưng toàn cục

1. MỞ ĐẦU

Phân loại hình ảnh là một trong những nhiệm vụ của lĩnh vực thị giác máy tính và học máy có giám sát, ở đó thuật toán tính toán và gắn nhãn cho hình ảnh từ một tập danh mục được xác định và đào tạo trước. Mục tiêu của bài toán này là xây dựng một mô hình học máy có khả năng học từ dữ liệu huấn luyện và sau đó sử dụng mô hình này để dự đoán lớp hoặc nhãn của các hình ảnh mới chưa từng xuất hiện trong quá trình học. Bước đầu tiên của quá trình đào tạo là trích xuất đặc trưng của các hình ảnh. Các đặc trưng này đại diện cho hình ảnh trong không gian chiều thấp và là cơ sở để thuật toán tiến hành phân loại. Dựa vào tính chất của việc phân loại, ta có thể chia bài toán phân loại hình ảnh thành 2 loại chính: phân loại nhãn đơn và phân loại đa nhãn. Phân loại nhãn đơn gán mỗi hình ảnh một nhãn duy nhất, có thể là phân loại đa lớp hoặc nhị phân. Phân loại đa nhãn cho phép mỗi hình ảnh có nhiều nhãn, phản ánh sự hiện diện của nhiều đối tượng hoặc thuộc tính khác nhau trong cùng một hình ảnh.

Phân loại hình ảnh đã trở thành một công nghệ quan trọng với nhiều ứng dụng đa dạng trong các lĩnh vực khác nhau. Đặc biệt trong y tế, nó đóng vai trò thiết yếu trong việc phát hiện và chẩn đoán các bệnh lý thông qua hình ảnh chụp từ MRI, CT scan, siêu âm hay kính hiển vi, giúp bác sĩ nhanh chóng xác định và điều trị bệnh. Trong ngành sản xuất, phân loại hình ảnh không chỉ kiểm tra chất lượng sản phẩm mà còn tự động hóa quy trình sản xuất bằng cách phát hiện và loại bỏ các sản phẩm lỗi, từ đó nâng cao chất lượng và giảm thiểu lãng phí. Tương tự, trong lĩnh vực an ninh, công nghệ này được sử dụng để nhận diện và giám sát các hoạt động đáng ngờ, như xâm nhập trái phép, giúp cải thiện khả năng giám sát và bảo vệ cộng đồng. Bên cạnh đó, trong giáo dục, phân loại hình ảnh hỗ trợ học tập và đào tạo bằng cách tự động nhận diện các đối tượng trong hình ảnh, tạo ra nội dung học tập tương tác và thú vị, giúp sinh viên và giáo viên hiểu rõ hơn về các khái niệm phức tạp.

Hơn nữa, trong thương mại điện tử, phân loại hình ảnh giúp tìm kiếm và tiếp thị sản phẩm hiệu quả hơn, tự động nhận diện sản phẩm và đề xuất các sản phẩm tương tự cho người dùng, nâng cao trải nghiệm mua sắm và tăng doanh số bán hàng. Nhờ những ứng dụng phong phú này, phân loại hình ảnh đang mở ra nhiều cơ hội và mang lại lợi ích vượt trội về hiệu suất, chất lượng công việc, và an toàn trong nhiều ngành nghề.

2. MỘT SỐ NGHIÊN CỨU LIÊN QUAN

Trong lĩnh vực phân loại hình ảnh, đã có nhiều nỗ lực nghiên cứu và phát triển các mô hình học máy để giải quyết bài toán này. Các phương pháp tiếp cận bắt đầu từ các kỹ thuật cổ điển, tập trung vào việc trích xuất các đặc trưng cục bộ. Đặc trưng cục bộ đóng vai trò quan trọng trong việc nhận diện các đối tượng và các chi tiết nhỏ trong hình ảnh. Một trong những kỹ thuật phổ biến nhất để trích xuất đặc trưng cục bộ là Scale-Invariant Feature Transform (SIFT), được phát triển bởi Lowe [1]. SIFT được tính toán từ cường độ của các vị trí xung quanh các điểm đặc biệt trong miền ảnh có thể được coi là điểm quan tâm hay còn gọi là keypoint. Các điểm này có đặc trưng là nơi mà hướng của đường biên thay đổi đột ngột hoặc là giao điểm của hai hay nhiều cạnh. Các keypoint có khả năng duy trì tính tỷ lệ khi hình ảnh được co giãn theo nhiều tỷ lệ khác nhau hay giữ nguyên tính chất của mình khi hình ảnh được xoay theo các góc độ khác nhau [8]. Điều này giúp cho các keypoint có thể được phát hiện và mô tả ở các tỷ lệ khác nhau trên cùng một hình ảnh một cách đáng tin cậy với mức độ lặp lại cao. Tuy nhiên, SIFT gặp hạn chế về tốc độ tính toán, khó áp dụng trong các ứng dụng yêu cầu xử lý thời gian thực.

Trong những năm gần đây, mạng nơ-ron tích chập (CNN) đã trở thành một công cụ mạnh mẽ và hiệu quả trong lĩnh vực phân loại hình ảnh [3]. CNN được thiết kế đặc biệt để xử lý dữ liệu dạng lưới như hình ảnh, nhờ khả năng tự động học và trích xuất

các đặc trưng quan trọng từ dữ liệu đầu vào thông qua các lớp tích chập, các lớp gộp và các lớp kết nối đầy đủ. Trong đó, các lớp tích chập đóng vai trò chính trong việc trích xuất các đặc trưng cục bộ từ hình ảnh bằng cách áp dụng các bộ lọc qua toàn bộ hình ảnh. Các lớp gộp giúp giảm kích thước không gian của đặc trưng, giữ lại các thông tin quan trọng và giảm bớt tính phức tạp của mô hình. Còn các lớp kết nối đầy đủ đóng vai trò kết hợp các đặc trưng đã trích xuất và thực hiện các phép phân loại cuối cùng, đưa ra dự đoán về nhãn của hình ảnh.

Visual Geometry Group (VGG) là một trong những mô hình CNN nổi bật, được phát triển bởi Simonyan và Zisserman [4] từ Đại học Oxford. VGG được biết đến với cấu trúc đơn giản nhưng hiệu quả, sử dụng các lớp tích chập với kích thước bộ lọc 3x3 và các lớp gộp 2x2. Mô hình này đã đạt được kết quả ấn tượng trên các bộ dữ liệu phân loại hình ảnh như ImageNet, và được sử dụng rộng rãi trong nhiều ứng dụng khác nhau. VGG có nhiều biến thể, phổ biến nhất là VGG-11, VGG-16 và VGG-19, với số lượng các lớp khác nhau. Với cấu trúc gồm 5 khối có cấu trúc tương tự nhau, biến đổi các ảnh có kích thước giảm dần và độ sâu tăng dần. Chính cấu trúc này của VGG giúp nó có khả năng trích xuất các đặc trưng toàn cục mạnh mẽ từ hình ảnh, giúp cải thiện hiệu suất phân loại và nhận dạng.

Bài báo này cung cấp một đánh giá toàn diện về hai phương pháp biểu diễn ảnh quan trọng: (1) Biểu diễn ảnh dựa trên đặc trưng cục bộ; (2) Biểu diễn ảnh dựa trên đặc trưng toàn cục. Trong đó, bài báo sẽ cung cấp cơ sở lý thuyết và thuật toán liên quan đến phân loại hình ảnh, bao gồm các mạng CNN đại diện cho phương pháp trích xuất đặc trưng toàn cục và các phương pháp truyền thống như SIFT đại diện cho phương pháp trích xuất đặc trưng cục bộ. Các thí nghiệm được thực hiện để đánh giá hiệu suất của hai phương pháp trên các tập dữ liệu thực tế cùng nhiều tiêu chí khác nhau. Kết quả thu được cung cấp một đánh giá toàn diện và khách quan về hiệu suất và ứng dụng thực tế của mỗi mô hình.

3. PHƯƠNG PHÁP NGHIÊN CỨU

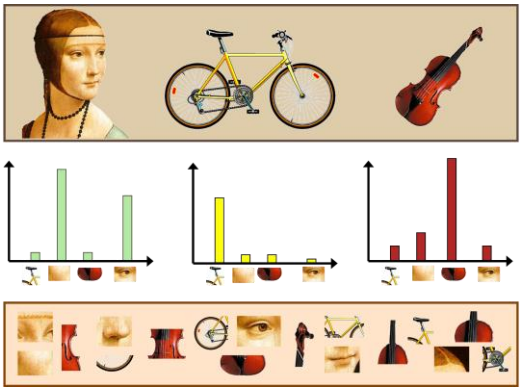
3.1. Biểu diễn hình ảnh với đặc trưng cục bộ

Trong phần này, phương pháp Bag of Visual Words (BoVW) [5] được sử dụng để biểu diễn hình ảnh dựa trên các đặc trưng cục bộ SIFT. BoVW, tương tự như Bag of Words trong xử lý ngôn ngữ tự nhiên, chuyển đổi hình ảnh thành một vector đặc trưng, trong đó mỗi phần tử biểu thị tần suất xuất hiện của một "từ" trực quan trong hình ảnh (visual word). Hình 1 minh họa cách BoVW biểu diễn hình ảnh dựa trên các đặc trưng cục bộ. Các bước thực hiện được mô tả chi tiết như sau:

(1) Trích xuất đặc trưng: Sử dụng SIFT để trích xuất các đặc trưng cục bộ từ mỗi hình ảnh.

(2) Xây dựng từ điển từ trực quan: Áp dụng thuật toán K-means để phân cụm các đặc trưng cục bộ trích xuất từ tập dữ liệu đào tạo, tạo thành từ điển từ trực quan với 200 cụm tương ứng với 200 từ trong từ điển.

(3) Mã hóa hình ảnh: Biểu diễn mỗi hình ảnh bằng một histogram của các từ trực quan dựa trên từ điển đã xây dựng.



Hình 1. Minh họa cách mô hình BoVW biểu diễn đặc trưng của hình ảnh

3.2. Biểu diễn hình ảnh với đặc trưng toàn cục

Trong quá trình thí nghiệm với mô hình VGG-11 gồm 11 lớp với dataset chó mèo, độ chính xác của mô hình chỉ rơi vào khoảng 50% trên toàn bộ tập dữ liệu và chỉ có thể nhận diện tốt được một trong hai đối tượng và không thể nhận diện được đối tượng còn lại. Vấn đề này xảy ra với cả hai bộ dataset chó mèo gồm 3000 mẫu (hình ảnh chưa xử lý) và 12000 mẫu (bao gồm hình ảnh đã xử lý). Bên cạnh đó, loss của mô hình hội tụ cũng rất nhanh chỉ sau 2 đến 3 epoch đầu và accuracy trên tập dữ liệu cũng hội tụ tại 0.5. Vì vậy, mạng VGG-8, một mạng mới gồm 8 lớp được trình bày như Bảng 1, được đề xuất để thực hiện các thí nghiệm. VGG-8 vẫn giữ nguyên cấu trúc 5 khối với kích thước ảnh đầu vào và đầu ra nhưng giảm bớt số lượng convolutional layer trong các khối và số lượng kernel trong mỗi layer. Tiếp đến, các lớp fully connected layer cũng được giảm xuống còn 4096 và 1024 đơn vị, lớp cuối cùng có số đơn vị tương ứng với từng tập dữ liệu được khảo sát. Việc giảm số lớp nhưng vẫn giữ nguyên cấu trúc và tính chất của mạng VGG cho hiệu suất tốt hơn mà vẫn giữ được các ưu điểm vốn có của mạng này.

Quá trình huấn luyện diễn ra với 30 epochs cùng batch size là 8 đảm bảo rằng mô hình có đủ thời gian để học các đặc trưng từ dữ liệu cũng như cân bằng giữa tốc độ huấn luyện và việc sử dụng tài nguyên tính toán với bộ dữ liệu lớn.

Bảng 1. Cấu trúc mạng VGG-8

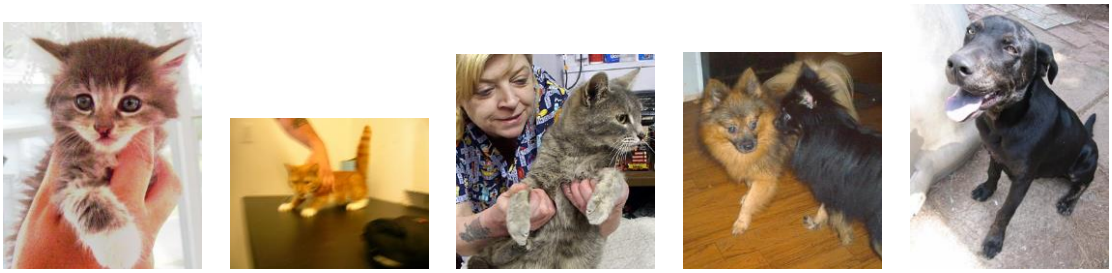
Input (224 pixel x 224 pixel RGB image)
Convolutional layer, 16 kernel 3 pixel x 3 pixel
Maxpool

Convolutional layer, 32 kernel 3 pixel x 3 pixel
Maxpool
Convolutional layer, 64 kernel 3 pixel x 3 pixel
Maxpool
Convolutional layer, 128 kernel 3 pixel x 3 pixel
Maxpool
Convolutional layer, 128 kernel 3 pixel x 3 pixel
Maxpool
Flatten
Fully Connected layer 4096 units
Fully Connected layer 1024 units
Fully Connected layer n units
Softmax

Trong kiến trúc VGG-8, hai lớp kết nối đầy đủ gồm 4096 units và 1024 units có thể được sử dụng để biểu diễn các vector đặc trưng. Do vậy, đặc trưng từ hai lớp kết nối này cũng được sử dụng để huấn luyện mô hình phân loại SVM. Thiết lập thí nghiệm này cung cấp đánh giá khách quan hơn về hiệu suất của hai phương pháp "trích xuất đặc trưng" là SIFT và VGG-8, mà không bị ảnh hưởng bởi độ chính xác của mạng neural đầy đủ.

4. KẾT QUẢ VÀ THẢO LUẬN

Thí nghiệm được thực hiện với hai bộ dữ liệu (xem Hình 2): bộ dữ liệu “chó mèo” được thu thập từ cuộc thi "Dogs vs. Cats Redux: Kernels Edition" trên nền tảng Kaggle với 1500 ảnh mỗi lớp chó mèo và bộ dữ liệu “ngôn ngữ ký hiệu số” do người dùng Muhammad Khalid chia sẻ trên nền tảng Kaggle với 1500 mẫu mỗi lớp từ 0 đến 9. Việc sử dụng hai tập dữ liệu để đánh giá giúp có cái nhìn khách quan hơn về hiệu suất của mô hình khi Tập dữ liệu chó mèo giúp đánh giá khả năng cơ bản của mô hình trong một môi trường đa dạng, trong khi tập dữ liệu ngôn ngữ ký hiệu số giúp kiểm tra tính ứng dụng và độ bền vững của mô hình trong các tình huống thực tế.



(a) Một số hình ảnh trong tập dữ liệu chó mèo

(b) Một số hình ảnh trong tập dữ liệu ngôn ngữ ký hiệu số

Hình 2. Minh họa các tập dữ liệu được sử dụng trong các thí nghiệm

Bên cạnh đó, thí nghiệm cũng bổ sung các biến thể ảnh để huấn luyện và đánh giá mô hình trên nhiều tiêu chí bao gồm: ảnh thường, ảnh lật, ảnh âm bản, ảnh bị bóp méo và ảnh xoay một góc bất kỳ (xem Hình 3). Trong đó, ảnh lật chỉ dùng khi đánh giá mô hình mà không tồn tại trong tập huấn luyện.

(c) Hình ảnh mèo trong tập dữ liệu chó mèo

(d) Hình ảnh số một trong tập dữ liệu ngôn ngữ ký hiệu số

Hình 3. Minh họa hình ảnh trong tập dữ liệu với các biến thể khác nhau (thứ tự từ trái qua phải: hình ảnh gốc (raw), hình ảnh âm bản (negative), hình đã bị lật (flipped), hình ảnh đã bị bóp méo (resized), hình ảnh bị xoay một góc bất kì (rotated)

Các mô hình được đánh giá bằng độ chính xác accuracy được tính bằng cách lấy số dự đoán đúng chia cho tổng số dự đoán. Accuracy đơn giản nhưng cung cấp cái nhìn tổng quan về hiệu suất của mô hình.

Dựa vào kết quả ở Bảng 2, dễ dàng nhận thấy với cả hai tập dữ liệu, mô hình VGG-8 có hiệu suất hẳn so với phương pháp SIFT. Thế nhưng mô hình VGG-8 lại gặp vấn đề overfitting nặng với tập dữ liệu chó mèo khi chênh lệch độ chính xác lên đến 12.61%, và có hiện tượng overfitting nhẹ với tập dữ liệu ngôn ngữ ký hiệu số. Ngược lại, tuy phương pháp SIFT khá ổn định ở cả hai tập dữ liệu khi chênh lệch độ chính xác chỉ xấp xỉ 2% nhưng lại gặp vấn đề underfitting khi độ chính xác thấp.

Điều này cho thấy SIFT hiệu quả cao trong việc nhận diện và phân loại hình ảnh trong nhiều bối cảnh khác nhau. Tuy nhiên phương pháp này lại gặp khó khăn với tập dữ liệu lớn khi nó không thể nắm bắt và tổng quát hóa các đặc trưng từ số lượng mẫu lớn, dẫn đến độ chính xác giảm sút. Hơn nữa, với bộ dữ liệu ngôn ngữ ký hiệu số, các đặc trưng cục bộ không đa dạng như với bộ dữ liệu chó mèo mà tương tự nhau như các góc ngón tay, ngoại cảnh đồng nhất. Từ đó, có thể thấy SIFT sẽ hoạt động tốt khi các đặc trưng cục bộ đủ đa dạng cùng số lượng mẫu vừa phải.

Bảng 2. Kết quả thí nghiệm trên các mô hình với tập kiểm thử và toàn bộ tập dữ liệu

Model	Testing set	Dataset	Difference
	Cat_Dog		
SIFT	70.11%	72.12%	2.01%
VGG-8	80.81%	93.42%	12.61%
4096-dims vector	80.86%	94.26%	13.40%
1024-dims vector	81.78%	94.53%	12.76%
	Hand_Sign		
SIFT	63.97%	65.68%	1.89%
VGG-8	90.31%	95.74%	5.43%
4096-dims vector	89.75%	96.83%	7.08%
1024-dims vector	89.48%	96.83%	7.35%

Ngược lại, VGG-8 lại làm việc khá hiệu quả khi số lượng dữ liệu huấn luyện lớn. Tập dữ liệu lớn và đa dạng giúp mô hình học được nhiều đặc trưng hơn và phát huy tối đã cấu trúc phức tạp của mạng tích chập, từ đó cải thiện khả năng tổng quát hóa và giảm nguy cơ overfitting. Từ đó cho thấy cần cân nhắc lựa chọn mô hình tùy vào độ lớn của tập huấn luyện.

Bên cạnh đó, khi sử dụng các vector đặc trưng được trích xuất từ mô hình kết quả xấp xỉ với khi chỉ sử dụng mô hình VGG-8 đơn thuần. Điều này cho thấy đã học được các đặc trưng tốt và các vector đặc trưng được trích xuất từ mô hình này vẫn giữ được những thông tin quan trọng, ngay cả khi kích thước vector được giảm đi. Và so với SIFT, mô hình SVM được huấn luyện bởi các vector này có độ chính xác cao hơn hẳn. Điều này cho thấy mạng CNN nói chung và mô hình VGG-8 nói riêng làm rất tốt trong việc trích xuất đặc trưng ảnh và cung cấp cho các mô hình phân loại khác những đặc điểm quan trọng cho đầu vào.

Bảng 3. Kết quả đánh giá các mô hình với nhiều tiêu chí

Criteria	SIFT	VGG-8	4096	1024
		Cat_Dog		
Raw	73.40%	97.27%	98.3%	98.17%
Negative	68.77%	90.80%	92.37%	92.07%
Resized	71.93%	97.57%	98.13%	98.07%
Rotated	74.37%	88.03%	89.33%	88.73%
Flipped	71.50%	76.90%	78.2%	76.9%
		Hand_Sign		
Raw	72.07%	99.46%	99.59%	99.53%
Negative	60.43%	98.54%	98.77%	98.89%
Resized	60.17%	99.38%	99.53%	99.55%
Rotated	70.78%	85.56%	89.43%	89.33%
Flipped	44.48%	46.59%	46.67%	46.71%

Bảng 3 thể hiện mức độ hiệu quả của phương pháp trích xuất đặc trưng SIFT và mô hình VGG-8 trên nhiều tiêu chí. Với phương pháp trích xuất đặc trưng SIFT, tập dữ liệu hình ảnh bị xoay một góc bất kì có độ chính xác gần nhất với tập dữ liệu gốc và có sự giảm nhẹ với tập dữ liệu hình ảnh bị bóp méo và ảnh âm. Tuy nhiên, với hình ảnh bị lật lại có sự biến động lớn khi với tập dữ liệu chó mèo thì độ chính xác không chênh lệch quá nhiều với các tiêu chí còn lại nhưng với tập dữ liệu ngôn ngữ ký hiệu số thì giảm đáng kể.

Với mô hình VGG-8, độ chính xác được duy trì trên ba tiêu chí: ảnh gốc, ảnh âm và ảnh bị bóp méo. Thế nhưng với ảnh xoay và ảnh bị lật, thí nghiệm với cả hai tập dataset đều chỉ rõ mô hình VGG-8 gặp khó khăn với hai tiêu chí này. Mô hình VGG-8

đạt độ chính xác cao trên các ảnh gốc và ảnh bị bóp méo, chứng tỏ rằng mô hình có khả năng học tốt từ các đặc trưng hình ảnh không thay đổi nhiều về không gian.

Sự giảm sút trên tiêu chí ảnh xoay hay ảnh lật của mô hình VGG-8 có thể được giải thích bởi phương pháp trích xuất đặc trưng toàn cục của nó: các mô hình CNN nói chung sử dụng giá trị pixel của ảnh để huấn luyện và dự đoán, không phải từ những đặc trưng cục bộ như SIFT. Do đó, khi hình ảnh bị biến đổi lớn về không gian như xoay hoặc lật, cấu trúc pixel thay đổi hoàn toàn, khiến cho mô hình không thể nhận diện được các đặc trưng đã học từ dữ liệu huấn luyện. CNN có thể học tốt từ các biến đổi nhỏ và cố định của hình ảnh, nhưng gặp khó khăn khi đối mặt với các biến đổi lớn làm thay đổi hoàn toàn cấu trúc hình ảnh. Kết quả này cho thấy việc chuẩn bị dữ liệu cho các mô hình CNN nói chung và VGG-8 nói riêng là rất quan trọng, vừa đảm bảo về số lượng dữ liệu cũng như sự đa dạng và phức tạp để mô hình có thể học một cách hiệu quả hơn.

Tương tự, kết quả này cho thấy độ chính xác của mô hình SVM khi huấn luyện bởi các vector đặc trưng được trích xuất bởi VGG-8 không quá chênh lệch so với mô hình VGG-8 đơn lẻ và vẫn giữ các đặc điểm chính của mô hình VGG-8. Các kết quả này cho thấy rằng việc sử dụng vector đặc trưng từ mô hình VGG-8 để huấn luyện mô hình SVM không mang lại sự cải thiện đáng kể về độ chính xác so với việc sử dụng mô hình VGG-8 đơn lẻ. Đồng thời, các đặc trưng học được từ mô hình VGG-8 vẫn giữ nguyên giá trị khi được sử dụng trong các tiêu chí khác nhau, chứng minh tính ổn định và khả năng tổng quát hóa của mô hình này.

5. KẾT LUẬN

Trong bài báo này, nhiều thí nghiệm đã được tiến hành để đánh giá hiệu suất của các phương pháp trích xuất đặc trưng và các mô hình phân loại hình ảnh khác nhau trên hai tập dữ liệu: chó mèo và ngôn ngữ ký hiệu số. Các kết quả phân tích và thảo luận cho thấy phương pháp biểu diễn đặc trưng cục bộ dựa trên SIFT cho khả năng phân loại ổn định trên các tập dữ liệu nhỏ và đa dạng như chó mèo, trong khi VGG-8, một CNN với cấu trúc sâu, đạt hiệu suất cao hơn đáng kể trên các tập dữ liệu lớn hơn như ngôn ngữ ký hiệu số. Bên cạnh đó, đánh giá về khả năng tổng quát hóa và hiện tượng overfitting, VGG-8 gặp vấn đề overfitting trên tập dữ liệu chó mèo, nhưng hoạt động tốt hơn trên tập dữ liệu ngôn ngữ ký hiệu số. Ngược lại, phương pháp biểu diễn đặc trưng cục bộ dựa trên SIFT không bị overfitting đáng kể, nhưng hiệu suất thấp hơn trên các tập dữ liệu lớn và ít đa dạng.

Như vậy, bài báo này đã cung cấp một cái nhìn toàn diện về hiệu suất của các phương pháp trích xuất đặc trưng và mô hình phân loại hình ảnh khác nhau. Những phát hiện này sẽ đóng góp quan trọng vào việc chọn lựa và phát triển các mô hình phân loại hình ảnh hiệu quả hơn trong nhiều ứng dụng.

TÀI LIỆU THAM KHẢO

[1]. D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision,* tập 60, số 2, pp. 91 -110, 2004.

[2]. T. Lindeberg, "Scale Invariant Feature Transform," 2012.

[3]. A. Krizhevsky, I. Sutskever và G. E. Hinton, "ImageNet Classification with Deep Convolutional Neural Networks," trong *Advances in Neural Information Processing Systems (NIPS)*, 2012.

[4]. K. Simonyan và A. Zisserman, "Very Deep Convolutional Networks for Large-Scale Image Recognition," 2014.

[5]. Sivic, and Zisserman. "Video Google: A text retrieval approach to object matching in videos." Proceedings ninth IEEE international conference on computer vision. IEEE, 2003.

COMPREHENSIVE UNDERSTANDING ON LOCAL AND GLOBAL FEATURE REPRESENTATIONS FOR IMAGE CLASSIFICATION

Huynh Van Nguyen Bao, Le Quang Chien

Faculty of Information Technology, University of Sciences, Hue University

Email: hvnguyenbao0611@gmail.com, lqchien@husc.edu.vn

ABSTRACT

Image classification is one of the important problems in the field of computer vision. This paper evaluates two main image classification methods: local and global features. Classical techniques such as SIFT focus on local features, while modern methods like Convolutional Neural Networks, exemplified by VGG-8, exploit global features. Experiments were conducted to compare the performance of these two methods on real-world datasets. The results provide a comprehensive assessment of each mothod's image classification capabilities and applications, which helps identify the strengths and weaknesses of each approach in different contexts.

Keywords: Computer vision, image classification, local features, global features.

Huỳnh Văn Nguyên Bảo Sinh ngày 06/11/2003 tại Thừa Thiên Huế. Hiện đang là sinh viên ngành Công nghệ thông tin, chuyên ngành Khoa học máy tính tai trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Thị giác máy tính, xử lý và phân loại ảnh, sinh ảnh tự động

Lê Quang Chiến Sinh ngày 15/09/1983 tại Thừa Thiên Huế. Năm 2005, ông tốt nghiệp cử nhân chuyên ngành Tin học tại trường Đại học Khoa học, Đại học Huế. Năm 2007, ông nhận bằng thạc sĩ chuyên ngành khoa học máy tính tại trường Đại học Khoa học, Đại học Huế. Năm 2016, ông nhận học vị tiến sĩ chuyên ngành Tin học tại trường SOKENDAI (The Graduate University for Advanced Studies), Nhật Bản. Hiện nay, ông đang công tác tại khoa Công nghệ Thông tin, trường Đại học Khoa học, Đại học Huế.

Lĩnh vực nghiên cứu: Xử lý và nhận dạng ảnh, xử lý video, học máy, thị giác máy tính. Ông hiện có một số công trình đăng ở các hội nghị và tạp chí khoa học quốc tế.