Xây dựng hệ thống phát hiện ảnh đồng phục từ mạng xã hội

Nguyễn Quang Mạnh, Nguyễn Đức Tuấn

***Tóm tắt*--Mạng xã hội đã tạo ra một số lượng lớn các hình ảnh theo thời gian vì các dữ liệu tải lên liên tục từ người sử dụng. Từ quan điểm về bảo mật kinh doanh, những người dùng mạng xã hội có thể vô tình tải lên các hình ảnh có chứa một số thông tin nguy hại. Các thông tin này có thể tiết lộ bí mật kinh doanh của một công ty nào đó. Trong bài báo này chúng tôi đề xuất một hệ thống phân tích nội dung ảnh cho bài toán phát hiện đồng phục từ dòng ảnh trên mạng xã hội. Hệ thống có thể giúp các nhà quản lý của công ty biết được khi nào đồng phục của công ty họ xuất hiện tên mạng xã hội. Bằng cách này, đối với những ảnh có chứa đồng phục của công ty, các nhà quản lý có thể xem xét các ảnh này xem nội dung của ảnh có chứa các thông tin gây ảnh hưởng đến công ty hay không và từ đó có những xử lý kịp thời. Thời gian xử lý và độ chính xác của hệ thống được đánh giá trên bốn bộ dữ liệu thực được lấy từ mạng xã hội.**

***Từ khóa*-- Học từ điển, Kim tự tháp không gian, Mã hóa thưa, Mạng xã hội, Phân loại ảnh.**

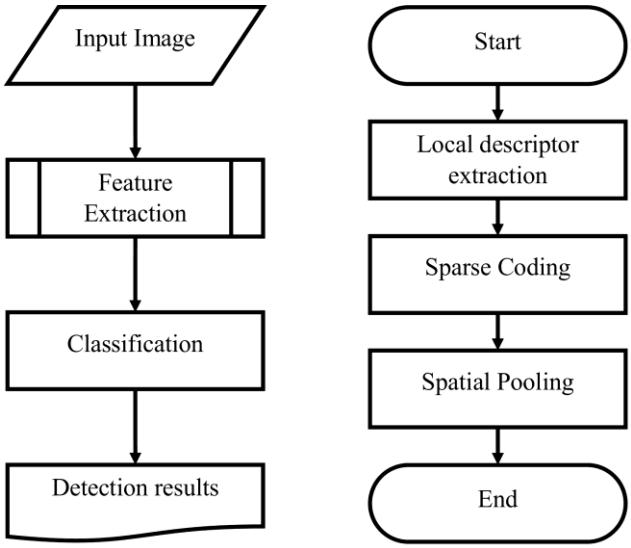
1. GIỚI THIỆU

Những năm gần đây chúng ta đã chứng kiến một sự bùng nổ nhanh chóng của mạng xã hội, trong đó các trang web chia sẻ hình ảnh như Facebook, Flickr, Photobucket cho phép người dùng tải lên các hình ảnh cá nhân của họ và hình ảnh họ chụp được. Nổi bật nhất là mạng xã hội Facebook hiện có số người dùng tương đương với dân số của các nước đông dân nhất thế giới, có khoảng 1,35 tỷ thành viên.

Một lượng thông tin khổng lồ tạo ra bởi các mạng xã hội đã kích thích con người khám phá, khai phá các kiến thức hữu ích, thông tin hữu ích được ẩn trong các dữ liệu mạng xã hội. Bên cạnh thông tin dưới dạng text như các câu cảm nhận, các bình luận, hoặc các thẻ tag, những bức ảnh trên mạng xã hội cũng có một nguồn thông tin giàu có tiềm ẩn bên trong. Việc khai thác các dữ liệu từ các bức ảnh này cũng đã thu hút sự tham gia của nhiều nhà nghiên cứu trên khắp thế giới. Rất nhiều công trình đã được đề xuất cho việc khai phá các thông tin từ các bức ảnh trên mạng xã hội dùng cả nội dung về hình ảnh và văn bản [11], [12], [13]. Cho đến bây giờ các hình ảnh trên mạng xã hội được khai thác nội dung với rất nhiều mục đích như khai thác thông tin về vị trí địa lý, phát hiện các sự kiện nóng của xã hội hoặc rất nhiều mục đích khác.

Nguyễn Đức Tuấn, sinh viên lớp Khoa học máy tính, khóa 55, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: newvalue92@gmail.com).

Nguyễn Quang Mạnh, sinh viên lớp KSTN-CNTT, khóa 55, Viện Công nghệ thông tin và Truyền thông, trường Đại học Bách Khoa Hà Nội (e-mail: nguyenquangmanh9099@gmail.com).



(a) Our framework **(b) Feature Extraction**

Hình 1: Mô hình chúng tôi đề xuất

Trong quá trình nghiên cứu, chúng tôi nhận thấy vấn đề phát hiện đồng phục của một công ty nào đó từ các hình ảnh trên mạng xã hội là một vấn đề hoàn toàn mới, chưa có các nghiên cứu nào thực hiện giải quyết bài toán này. Bài toán phát hiện đồng phục có thể phục vụ cho mục đích giám sát nhân viên, quản lý hình ảnh và thông tin của công ty. Trong trường hợp thứ nhất, nếu nhân viên trong giờ làm việc tự ý đi mua sắm thì công ty có thể phát hiện ra và cần có biện pháp xử lý. Một số tổ chức đòi hỏi nhân viên phải giữ tác phong và có cách hành xử đúng mực ở nơi làm việc và nơi công cộng, việc phát hiện ra nhân viên xuất hiện ở những nơi không phù hợp có thể ảnh hưởng đến hình ảnh và uy tín của cơ quan, tổ chức. Trong một trường hợp khác, nếu nhân viên bị phát hiện gặp gỡ với đối thủ thì cũng là một tín hiệu cảnh báo đối với khả năng bị lộ thông tin, làm ảnh hưởng đến sự rò rỉ thông tin nội bộ. Trong nghiên cứu này, chúng tôi xây dựng một hệ thống thu thập dữ liệu để liên tục lấy hình ảnh từ mạng xã hội, nó hình thành nên một dòng dữ liệu, sau đó chúng tôi đề xuất một hệ thống phân tích ảnh sử dụng các công cụ tiên tiến để phát hiện và lọc ra hình ảnh chứa trang phục từ dòng dữ liệu này.

Một ứng dụng quan trọng khác của việc khai phá thông tin từ hình ảnh trên các mạng xã hội là giải quyết vấn đề tự động gán thẻ tag cho hình ảnh. Bài toán gán thẻ tự động là một phần quan trọng của hệ thống tìm kiếm hình ảnh. Hầu hết các hệ thống lớn đảm nhiệm việc lưu trữ hình ảnh đều sắp xếp và tổ chức các hình ảnh trong cơ sở dữ liệu của họ dựa trên các lời chú thích của các

bức ảnh. Các lời chú thích này được gán bằng tay tùy vào nội dung của ảnh. Vì thế việc giải quyết bài toán gán thẻ tag một cách tự động cho các hình ảnh là cần thiết. Bài toán này sẽ giúp chúng ta giảm chi phí và thời gian khi thực hiện viết lời chú thích cho từng bức ảnh chưa được gán thẻ tag. Hệ thống dùng cho việc phát hiện đồng phục có thể được dùng để gán các từ khóa cho hình ảnh ví dụ như các từ khóa về tên công ty, vị trí và nhiều từ khóa quan trọng khác.

Ngày nay có nhiều hệ thống giới thiệu và các dịch vụ quảng cáo trực tuyến khai thác các nội dung dạng văn bản được học, được trích xuất từ các hồ sơ của người dùng. Với các dữ liệu thu được, các hệ thống và các dịch vụ quảng cáo có thể gợi ý các mặt hàng, các nội dung quảng cáo được thiết kế cho từng người dùng riêng biệt, phù hợp với thị hiếu của họ. Nó nảy sinh ra một câu hỏi thú vị là liệu chúng ta có thể xây dựng các hệ thống tương tự như khai thác dữ liệu hình ảnh cá nhân, các hình ảnh này thường chứa nhiều dữ liệu ẩn hơn so với các dữ liệu dạng văn bản được lấy từ hồ sơ người dùng. Khai thác hình ảnh cá nhân có thể giúp chúng ta tìm hiểu các sở thích của người dùng cá nhân, có thể được sử dụng cho xây dựng các hệ thống giới thiệu, các dịch vụ quảng cáo tốt. Một hệ thống cho phát hiện đồng phục có thể được tích hợp như một hệ thống gợi ý. Ví dụ, khi người dùng xem các hình ảnh có chứa đồng phục của một công ty, chúng ta cũng có thể tiên đoán rằng người dùng có thể quan tâm đến công ty đó. Do đó, các thông tin về công ty hoặc các sản phẩm của công ty đó có thể được đẩy lên nội dung quảng cáo.

Phát hiện đồng phục là một trường hợp đặc biệt của bài toán phân loại hình ảnh. Cụ thể hơn, đây là một vấn đề phân loại hai lớp. Một hệ thống tổng quát cho huấn luyện bộ phân loại hình ảnh bao gồm các bước: Thu thập dữ liệu, trích chọn đặc trưng ảnh, và học mô hình phân loại. Trong quá trình thực hiện phân loại, sau khi trích chọn đặc trưng, các hình ảnh sẽ được phân loại bởi bộ phân loại đã được học. Từ các dữ liệu thu về bằng dòng dữ liệu trên mạng xã hội, một hệ thống cho phát hiện đồng phục cần làm việc đủ nhanh để phù hợp với tốc độ tải xuống hình ảnh trong khi vẫn đảm bảo độ chính xác phát hiện đồng phục. Trong bài báo này chúng tôi đề xuất một hệ thống hoạt động thời gian thực cho việc phát hiện đồng phục có thể xử lý lên đến 40 ảnh trên giây với kích thước bình thường.

2. NGHIÊN CỨU LIÊN QUAN

Có rất nhiều nguồn thông tin hữu ích có thể được khai thác từ các hình ảnh trên mạng xã hội. Trong [2] Crandall đã đề xuất một cách tiếp cận để tổ chức 35 triệu hình ảnh thu được từ Flickr sử dụng phân tích nội dung dựa trên văn bản và nội dung hình ảnh với cấu trúc của mô hình phân tích dựa trên các dữ liệu về địa lý. Trong [6] Luo đã kết hợp các thông tin từ vệ tinh với các nội dung hình ảnh để nhận dạng môi trường mà ảnh được chụp. Trong [1] Chen đã khai thác các dữ liệu dưới dạng siêu dữ liệu của các hình ảnh từ Flickr bao gồm các thông tin về thời gian, thông tin về địa điểm và thẻ tag người dùng định nghĩa để phân tích sự phân bố của các hình ảnh và tự động phát hiện các sự kiện. Gắn thẻ tag tự động là một nhiệm vụ khó khăn của học máy nhưng nó là một bài toán con rất quan trọng cho nhiều ứng dụng. Phương pháp thông thường để xử lý với bài toán này là dùng học máy học được mô hình phân loại, ví dụ SVM [3] học mô hình phân loại từ bộ dữ liệu học được gán nhãn bằng tay với một tập các từ khóa xác định trước. Một cách tiếp cận đầy hứa hẹn khác là mô hình tìm kiếm dựa trên đánh chỉ mục nhanh và các kỹ thuật tìm kiếm như các hàm băm [8] hoặc tìm kiếm theo đối sánh phong cảnh [7]. Gần đây, trong [12] Pengcheng We đã khảo sát một phương pháp mới đó là tự động gán thẻ tag cho ảnh thông qua việc học một độ đo khoảng cách hiệu quả dựa trên các nội dung về văn bản và hình ảnh. Trong [13] Yu đã đề xuất một phương pháp tự động gợi ý các nhóm hình ảnh dựa trên bộ sưu tập của từng cá nhân.

Sự thành công của một hệ thống cho bài toán phân loại ảnh phụ thuộc rất lớn vào việc trích xuất đặc trưng từ ảnh. Trong [14] Oliva đã đề xuất một cách biểu diễn cho các ảnh phong cảnh bằng cách dùng GIST Descriptor. Tuy nhiên, GIST chỉ biểu diễn một quan sát tổng quát mà bỏ qua các chi tiết đối tượng trong cảnh, điều này làm cho GIST chỉ phù hợp với các hình ảnh ngoài trời. Trong [15] Wu và Rehg đã đề xuất Centrist Descriptor cái mà kết hợp cả các thông tin địa phương và thông tin toàn cục của ảnh. Một điểm yếu của Centrist Descriptor là không bất biến với các phép quay. Trong [16] Lowel đã đề xuất SIFT Descriptor, bất biến với các phép biến đổi như phép quay, phép biến hình thông qua thay đổi kích thước, phép chuyển đổi Affine và dưới các điều kiện chiếu sáng khác nhau. Do đó, SIFT Descriptor đã được dùng phổ biến trong nhiều lĩnh vực của thị giác máy tính trong đó bao gồm cả bài toán phân loại ảnh. Một đặc trưng cải tiến của SIFT Descriptor là Dense SIFT Descriptor (DSIFT Descriptor) được đề xuất bởi Bosch [17]. Thay vì cách tính SIFT Descriptor ở các điểm Interest Points, DSIFT Descriptor được tính tại các điểm trên lưới của bức ảnh. Bằng cách làm này, DSIFT Descriptor sẽ thu được nhiều thông tin hơn và hướng tới việc có kết quả tốt hơn SIFT Descriptor với những ứng dụng về phân loại ảnh. Trong [18] Bay đã đề xuất SURF Descriptor bằng cách dựa vào biến đổi Wavelts (Hàm Haar). So sánh với SIFT, SURF làm việc nhanh hơn, tuy nhiên lại cho độ chính xác thấp hơn SIFT. Trong hệ thống chúng tôi đề xuất dùng cho bài toán phát hiện đồng phục, chúng tôi dùng DSIFT Descriptor để thực hiện trích xuất đặc trưng từ ảnh.

Trong hệ thống được nghiên cứu, chúng tôi đề xuất dùng bộ mô tả DSIFT trong mô hình biểu diễn mã thưa. Sau khi đặc trưng DSIFT của ảnh được trích xuất, các đặc trưng này sẽ được chuyển thành một đặc trưng ở cấp độ cao bằng các phương pháp mã hóa thưa (Sparse coding) và kim tự tháp không gian (Spatial pyramid). Mã hóa thưa là một phương pháp hiệu quả được áp dụng để giải quyết rất nhiều bài toán trong thị giác máy tính bao gồm cả phân loại ảnh. Với bài toán phân loại ảnh, đã có nhiều phương pháp học từ điển (từ điển phục vụ cho mã hóa thưa) mới với hiệu quả cao. Thực hiện học một từ điển over-complete đóng một vai trò quan trọng trong sự thành công của bài toán phân loại dùng mã hóa thưa. Chúng ta mong muốn xây dựng được một từ điển vừa tốt cho việc phân loại, vừa tốt cho việc biểu diễn, có thể là biểu diễn tín hiệu hoặc biểu diễn ảnh. Dựa vào KSVD [19], Zhang và Li [10] đã đề xuất một phương pháp có tên là Discriminative KSVD (DKSVD) với việc đưa thêm thông tin về độ lỗi của bộ phân loại vào hàm mục tiêu của

KSVD. Jiang [5] đã đề xuất một phương pháp mới gọi là LC-KSVD bằng việc bổ sung thêm thông tin ràng buộc về nhãn của dữ liệu học. Để cải thiện ràng buộc về tính phân biệt giữa các từ trong từ điển, Yang [9] đã đề xuất FDDL(Fisher Dictionary Learning) dùng tiêu chuẩn Fisher để làm cho hệ số biểu diễn thưa phân biệt hơn trên từng lớp dữ liệu.

Thông thường, trong hầu hết các phương pháp học từ điển cho mã hóa thưa được đề xuất từ trước đến giờ đều dùng các chuẩn l_0 và l_1 để ràng buộc hệ số biểu diễn sau mã hóa phải thưa. Tuy nhiên, những chuẩn này dẫn tới việc chúng ta cần nhiều thời gian trong quá trình học và quá trình test. Gần đây, trong một vài phương pháp học từ điển với độ chính xác cao đã dùng sang chuẩn l_2 để làm ràng buộc thưa cho hệ số biểu diễn của mã hóa thưa, cụ thể trong [4] và [20]. So sánh với các phương pháp trước đó, các phương pháp học từ điển dùng chuẩn l_2 không những giảm về thời gian tính toán mà còn cho độ chính xác ấn tượng đối với bài toán phân loại hình ảnh. Do đó chuẩn l_2 là một lựa chọn tốt cho hệ thống của chúng tôi để đảm bảo về thời gian tính toán và vẫn đảm bảo về độ chính xác của bài toán phát hiện đồng phục.

3. HỆ THỐNG ĐỀ XUẤT

Hệ thống của chúng tôi cho việc phát hiện đồng phục với thời gian thực được miêu tả ở Hình 1. Vấn đề phát hiện đồng phục trong hệ thống này chính là bài toán phân loại ảnh. Mỗi ảnh từ dòng ảnh (image stream) sẽ được chia thành hai lớp: đồng phục và không đồng phục. Với mỗi hình ảnh, chúng tôi thực hiện trích xuất đặc trưng, sau khi có vector đặc trưng cho ảnh, vector này sẽ là đầu vào cho mô hình học từ bộ dữ liệu học đã được thực hiện trước.

Quá trình trích xuất đặc trưng bao gồm ba bước: trích xuất các descriptor địa phương, mã hóa thưa và biểu diễn đặc trưng dùng kim tự tháp không gian.

A. Trích xuất đặc trưng

1) Descriptor địa phương

Trích xuất đặc trưng của ảnh liên quan đến việc giảm số lượng dữ liệu phải xử lý, thay vì chúng ta phải xử lý một tập dữ liệu lớn, chúng ta thực hiện trích xuất đặc trưng để khối lượng tính toán nhẹ hơn mà vẫn đảm bảo độ chính xác cần thiết, bước này đặc biệt quan trọng với các ứng dụng thời gian thực. Ngoài ra, việc trích xuất đặc trưng cũng giúp tìm ra các đặc trưng tiềm ẩn trong dữ liệu, điều này giúp làm tăng độ chính xác của bài toán. Trong xử lý ảnh, đã có nhiều phương pháp được nghiên cứu để trích xuất đặc trưng từ ảnh. Chúng tôi sẽ xem xét một vài đặc trưng điển hình như SIFT, SURF, Text-on Boost. Mỗi đặc trưng đều có những ưu và nhược điểm riêng.

SIFT trích xuất một tập các đặc trưng từ ảnh, điểm mạnh của SIFT là bất biến đối với các phép quay và phép thay đổi tỉ lệ. SIFT cho độ chính xác khá ấn tượng cho rất nhiều bài toán trong thị giác máy tính. Tuy nhiên do nhược điểm về thời gian tính toán lâu nên đặc trưng SIFT không phù hợp cho bài toán phát hiện đồng phục.

SURF có cách thực hiện giống như SIFT, tuy nhiên về thời gian tính, SURF thực hiện nhanh hơn do việc dùng hàm Haar của biến đổi wavelet để trích xuất các điểm đặc trưng. Tuy có ưu điểm về thời gian nhưng SURF lại có nhược điểm về độ chính xác. Do đó, SURF không phù hợp với ứng dụng thời gian thực.

Text-on Boost cũng là một đặc trưng mạnh, tuy nhiên về thời gian, đặc trưng này cũng không thể đáp ứng cho các ứng dụng thời gian thực.

Tóm lại, khi trích xuất đặc trưng từ một ảnh, chúng tôi phải cân nhắc giữa việc lựa chọn đặc trưng để vừa đảm bảo về độ chính xác, vừa phải đáp ứng được yêu cầu về thời gian để đảm bảo xử lý của luồng ảnh trực tuyến tải xuống. Để thu được tập đặc trưng từ dữ liệu, chúng tôi quyết định dùng đặc trưng DSIFT, DSIFT vừa đảm bảo về mặt độ chính xác, bên cạnh đó cũng có thời gian tính ở mức chấp nhận được. Sau quá trình thử nghiệm chúng tôi nhận thấy DSIFT phù hợp cho bài toán phát hiện đồng phục cả về thời gian và độ chính xác, phù hợp với việc xử lý dòng ảnh trực tuyến (image stream).

2) Mã hóa thưa

Mã hóa thưa đã được nghiên cứng rộng rãi và được áp dụng trong nhiều bài toán của xử lý ảnh, mã hóa thưa cũng đạt được nhiều kết quả ấn tượng cho nhiều bài toán phân loại ảnh. Chúng tôi dùng phương pháp này để biểu diễn đặc trưng DSIFT sau quá trình trích xuất đặc trưng từ ảnh. Sau khi thực hiện quá trình trích xuất đặc trưng DSIFT từ tập ảnh, ta được tập đặc trưng X (đây chính là ma trận, mỗi cột biểu thị cho một vector đặc trưng trích tại một điểm trên lưới (lưới DSIFT). Chúng tôi đã thử nghiệm một số phương pháp học từ điển như KSVD, MI-KSVD để thu được từ điển biểu diễn cho mỗi vector đặc trưng (mỗi cột của X).

KSVD là một phương pháp tổng quá của K-means. Hàm mục tiêu của KSVD như sau:

$$\min_{D,A}\left\{\left\|X-DA\right\|_F^2\right\}\ \ s.t.\ \forall i,\left\|a_i\right\|_0\le T$$

Trong đó A là ma trận hệ số của ma trận đặc trưng X , T là độ thưa của vector hệ số biểu diễn. Hiệu năng của KSVD là khá tốt, tuy nhiên chúng tôi nhận thấy từ điển D được học bởi KSVD vẫn chưa đủ tốt. Tất cả các từ trong từ điển D vẫn khá tự do, trong hàm mục tiêu của KSVD chưa có ràng buộc cho các từ này. Do đó, chúng tôi quan tâm đến mối quan hệ giữa các từ trong từ điển (cụ thể là sự tương quan giữa các từ). Đó là ý tưởng của MI-KSVD (Mutual Incoherent-KSVD). Hàm mục tiêu của MI-KSVD như sau:

$$\min_{D,A}\left\{\left\|X-DA\right\|_F^2\right\}+\lambda\sum_{i=1}^{M}\sum_{j=1,j\ne i}^{M}\left|d_i^Td_j\right|$$
$$s.t.\forall m,\left\|d_m\right\|_2=1\,and\ \forall n,\left\|x_n\right\|_0\le T$$

Trong đó M là số lượng từ trong từ điển D , $\lambda\ge 0$ là tham số điều chỉnh sự đóng góp của ràng buộc về sự tương quan giữa các từ.

Tóm lại, chúng tôi dùng từ điển D để biểu diễn cho các đặc trưng sau khi dùng DSIFT trích xuất từ ảnh. Mặc dù mã hóa thưa bao gồm cả học có giám sát và không giám sát, tuy nhiên đối với bài toán này thì việc học không giám sát cho kết quả cao hơn. Chúng tôi gợi ý việc dùng học không giám sát vì việc gán nhãn cho các đặc trưng sẽ gây ra lỗi dữ liệu và điều này làm giải độ chính xác cho bài toán phân loại.

3) Kim tự tháp không gian

Các hình ảnh được tải về từ mạng Internet rất đang dạng về kích thước, do vậy khi trích xuất đặc trưng DSIFT, chúng ta sẽ thu được tập các vector đặc trưng trên mỗi ảnh có số lượng khác nhau. Mỗi vector đặc trưng sẽ được biểu diễn bằng mã hóa thưa. Vấn đề đặt ra là cần chuẩn để trên mỗi ảnh chúng ta sẽ thu được một vector đặc trưng với số chiều giống nhau. Phương pháp kim tự tháp không gian sẽ cho phép làm điều này. Dựa vào vị trí của từ vector đặc trưng DSIFT, phương pháp kim tự tháp không gian sẽ thực hiện gom nhóm các vector này và thực hiện chọn ra một vector mới đại diện cho cụm đó. Số cụm sẽ được xác định tùy vào cách chia. Ví dụ, nếu chúng ta chọn cấp độ là 2, chúng ta sẽ có 4 cụm các vector đặc trưng. Trên một cụm, chúng ta có thể lấy vector mới theo cách chọn giá trị lớn nhất hoặc trung bình. Chúng tôi đã khảo sát và nhận thấy việc chọn giá trị lớn nhất sẽ cho độ chính xác cao hơn. Thông thường chúng tôi sẽ chọn kim tự tháp với 3 mức, mức thứ nhất có cấp độ là 4, mức thứ hai cấp độ là 2, mức thứ ba cấp độ là 1. Như vậy chúng tôi sẽ có tổng số cụm là 21 cụm. Khi lấy giá trị lớn nhất và làm theo cách như trên, chúng ta sẽ thu được một vector đặc trưng với số chiều giống nhau trên các bức ảnh, ngay cả khi các bức ảnh này có kích thước khác nhau. Cách làm này mang lại độ chính xác khá cao cho bài toán phát hiện đồng phục đồng thời cũng khắc phục được tình trạng khác nhau về kích thước của hình ảnh.

B. Phân loại

Để đáp ứng được yêu cầu thời gian thực của hệ thống, chúng tôi cần một bộ phân loại không chỉ mạnh mà còn phải phân loại nhanh chóng. Sau khi cân nhắc kỹ càng, chúng tôi chọn DPL để sử dụng trong hệ thống của chúng tôi.

DPL đang là một bộ phân loại mới nhất và mạnh mẽ theo dựa trên tiếp cận học từ điển. Khác với các cách tiếp cận học từ điển khác cố gắng xây dựng một từ điển duy nhất cho cả việc biểu diễn và phân loại như trong DKSVD [10], LCKSVD [5], FDDL [9], DPL tách biệt hai chức năng này thông qua việc học hai từ điển riêng việt một từ điển dung cho việc biểu diễn tốt dư liệu gọi là từ điển tổng hợp và một từ điển cho việc phân loại gọi là từ điển phân tích. Hàm mục tiêu của DPL được biểu diễn bởi công thức:

$$\left\{P^*, D^*\right\} = \underset{D,K}{\operatorname{argmin}} \sum_{k=1}^{K} \| X_k - D_k P_k X_k \|_F^2 + \lambda \| P_k * \overline{X}_k \|_F^2 \text{ s.t} \| d_i \|_F^2 \le 1 \quad (1)$$

Trong công thức **(1)** , P từ điển phân tích, P có cấu trúc $P = [P_1; ..; P_k; ..; P_K]$, D là từ điển tổng hợp, D có cấu trúc $D = [D_1, .., D_k, .., D_K]$, mối cặp $\{D_k, P_k\}$ là cặp từ điển con tương ứng với lớp k. X_k biểu diễn dữ liệu thuộc lớp k, và $\overline{X}_k$ biểu diễn dữ liệu không thuộc lớp k, d_j biểu diễn từ thứ j trong từ diển D. Trong mô hình này, tính phân biệt của mô hình được tạo bởi các cấp từ điển con. Trong khi P_k giúp đảm bảo hệ số biểu diễn của từ điển P_k cho dữ liệu thuộc lớp khác rất gần 0, D_k đảm bảo từ điển thuộc mỗi lớp biểu diễn tốt dữ liệu thuộc chính lớp đó.

Bài toán phát biển bởi công thức không là hàm lồi. Để giải quyết bài toán đó, nhóm tác giả đã nới lỏng điều kiện để đưa về bài toán sau:

$$\left\{P^*, D^*, A^*\right\} = \underset{\text{P,D}}{\operatorname{argmin}} \sum_{k=1}^{K} (\|X_k - D_k A_k\|_F^2 + \tau \|P_k * X_k - A_k\|_F^2 + \lambda \|P_k * \overline{X}_k\|_F^2) \text{ s.t. } \| d_j \|_F^2 \le 1 \quad (2)$$

Quá trình tối ưu hàm mục tiêu **(2)** được tiến hành thông qua việc liên tục giải quyết qua 2 vấn đề:

(1) Cố định D và P, cập nhật A

$$A^* = \underset{A}{\operatorname{argmin}} \sum_{k=1}^{K} \|X_k - D_k A_k\|_F^2 + \tau \|P_k * X_k - A_k\|_F^2 \quad (3)$$

Bài toán trên có lời giải tối ưu dạng:

$$A^* = \left(D_k^T D_k + \tau I\right)^{-1} * \left(\tau * P_k * X_k + D_k^T X_k\right) \quad (4)$$

(2) Cố định A, cập nhật D và P

$$\begin{cases} P^* = \underset{P}{\operatorname{argmin}} \sum_{k=1}^{K} \|P_k * X_k - A_k\|_F^2 + \lambda \left\|P_k * \overline{X}_k\right\| \\ D^* = \underset{D}{\operatorname{argmin}} \sum_{k=1}^{K} \|X_k - D_k * A_k\|_F^2 \text{ s.t. } \|d_i\|_F^2 \le 1 \end{cases} \quad (5)$$

Lời giải P có thể nhận được theo công thức:

$$P_k^* = \tau A_k X_k \left(\tau X_k X_k^T + \lambda \overline{X_k}\,\overline{X_k^T}\right)^{-1} \quad (6)$$

Trong đó $\lambda = 10e^{-4}$ là một số nhỏ. Để tìm D, ta sử dụng thêm biến S:

$$\min_{D,S} \sum_{k=1}^{K} \|X_k - D_k A_k\| \text{ s.t. } D = S, \|s_i\|_F^2 \le 1 \quad (7)$$

Bài toán trên có thể được giải bằng giải thuật ADMM:

$$\begin{cases} D^{(r+1)} = \underset{D}{\operatorname{argmin}} \sum_{k=1}^{K} \|X_k - D_k A_k\|_F^2 + \rho \left\|D_k - S_k^{(r)} + T_k^{(r)}\right\|_F^2 \\ S^{(k+1)} = \underset{S}{\operatorname{argmin}} \sum_{k=1}^{K} \rho \left\|D_k - S_k^{(r+1)} + T_k^{(r+1)}\right\|_F^2 \\ T^{(r+1)} = T^{(r)} + D_k^{(r+1)} - S_k^{(r+1)}, \text{ update } \rho \text{ if appropriate} \end{cases} \quad (8)$$

Phép phân loại trong mô hình DPL rất đơn giản, dựa trên lỗi biểu diễn. Cho mẫu y, nhãn của mẫu được xác định theo công thức:

$$identity(y) = \underset{i}{\operatorname{argmin}} \left\|y - D_i * P_i * y\right\|_F^2 \quad (9)$$

Qua công thức, phép phân loại của DPL chỉ cần thực hiện phép nhân ma trận. Do vậy, thời gian tính toán của DPL rất nhanh. Kết quả thực nghiệm và thời gian tính của DPL được trình bày chi tiết trong [6].

4. THỰC NGHIỆM

A. Các bộ dữ liệu

Ngày nay mạng xã hội đã trở thành phổ biến, ngày càng nhiều thông tin được chia sẻ hơn, các thông tin có thể dưới nhiều dạng, dạng văn bản, dạng hình ảnh. Các nhân viên của nhiều công ty cũng như vậy, họ có thể chia sẻ các hình ảnh tự chụp về bản thân họ hoặc về công ty của họ; các người dùng khác cũng có thể tải lên các hình ảnh họ chụp được. Nắm bắt được tình trạng này, chúng thôi đã thực hiện tải xuống các hình ảnh trên mạng xã hội và phân loại để phát hiện xem liệu các nhân viên của một công ty có chia sẻ các hình ảnh chứa đồng phục công ty của họ hay không. Trong nghiên cứu này chúng tôi tập trung vào mạng xã

hội Facebook và máy tìm kiếm Google.

Với mạng xã hội Facebook, có thể nói đây là mạng xã hội phổ biến nhất với lượng người dùng rất lớn, chúng tôi thực hiện tải xuống các hình ảnh từ tập các album được công khai trên các fan-page. Bằng cách làm đó, chúng tôi không vi phạm bản quyền hoặc các quyền riêng tư của từng cá nhân. Chúng tôi nhận thấy các hình ảnh thu được rất đa dạng, nhiều chủ đề, nhiều kích thước, đa màu sắc và ánh sáng. Chủ đề của các hình ảnh có thể là về phong cảnh, đường phố, ảnh về các cuộc dã ngoại hoặc ảnh tự sướng (tự chụp về bản thân) của mọi người. Kích thước của các hình ảnh cũng đa dạng, có rất nhiều ảnh với kích thước lớn được chụp từ các máy ảnh (kích thước khoảng 1900x1700) hoặc cũng có những ảnh nhỏ được chụp từ điện thoại… Màu sắc và độ sáng của các ảnh cũng rất khác nhau, các ảnh tự chụp (selfie) có thể được chụp dưới điều kiện thiếu sáng hoặc phơi sáng rất cao. Bên cạnh đó, có thể có một hoặc nhiều đối tượng trong ảnh. Tất cả các đặc điểm về ảnh này sẽ là thách thức cho bài toán phân loại hình ảnh. Để tải xuống các album ảnh, chúng tôi đã dùng mô hình đồ thị của Facebook (Facebook Graph), mô hình này được Facebook cung cấp sẵn thông qua các API, chúng tôi dùng Python gửi yêu cầu đến Facebook dùng các API, sau đó sẽ nhận được các đường link đến từng ảnh, sau đó chúng tôi thực hiện tải xuống các hình ảnh từ tập các liên kết (link) và lưu lại vào một thư mục để thực hiện phục vụ cho bài toán phân loại. Để xây dựng dữ liệu cho thử nghiệm, hiện tại Facebook chưa có công cụ tìm kiếm ảnh giống nhau theo chủ đề hoặc nội dung nên chúng tôi dùng tập các ảnh lấy từ Facebook cho vào lớp không có đồng phục.

Chúng tôi dùng công cụ tìm kiếm Google để xây dựng bộ ảnh đồng phục. Chúng tôi thực hiện truy vấn các đoạn văn bản có liên quan đến tên đồng phục đến máy tìm kiếm Google, kết quả trả về của Google sẽ là tập các liên kết đến các hình ảnh, những hình ảnh này có thể có liên quan tới nội dung đoạn văn bản đã được truy vấn. Phần lớn các hình ảnh từ Google trả về đều phù hợp với đoạn văn bản đã truy vấn, chúng tôi cũng thực hiện một bước tiền xử lý để loại bỏ những mẫu sai trước khi xây dựng dữ liệu cho mô hình. Từ tập các liên kết trả về, chúng tôi thực hiện tải xuống các ảnh này và lưu lại vào thư mục. Mặc dù các hình ảnh tải xuống khá phù hợp với trang phục chúng tôi cần, tuy nhiên cũng giống như các hình ảnh lấy từ mạng xã hội, các hình ảnh của Google cũng rất đa dạng, giống như các đặc điểm đã phân tích ở phần ảnh lấy từ Facebook. Điều này đặt ra khá nhiều thách thức cho bộ phân loại. Từ cách ảnh này chúng tôi thực hiện xây dựng tập ảnh chứa dữ liệu đồng phục mà chúng tôi quan tâm để làm dữ liệu học và đánh giá cho mô hình.

Từ các mẫu dữ liệu, các hình ảnh thu thập được, chúng tôi thực hiện đánh giá độ chính xác của mô hình. Chúng tôi đã thực hiện tải xuống bốn bộ dữ liệu. Mỗi bộ dữ liệu có nội dung về một bộ trang phục khác nhau. **Bảng 1** sẽ chỉ ra chi tiết các bộ dữ liệu. Bộ Lawson là bộ đồng phục của công ty Lawson, các bộ còn lại là các bộ ảnh về đồng phục thi đấu bóng đá của các câu lạc bộ lớn.

Bảng 1: Chi tiết về các bộ dữ liệu ảnh đồng phục

STT	Tên bộ dữ liệu	Số ảnh học	Số ảnh đánh giá
1	Lawson	190	245
2	Argentina	195	250
3	Brazil	195	250
4	Barcelona	327	394

B. Môi trường thử nghiệm

Chúng tôi đã thực hiện cài đặt hệ thống phát hiện đồng phục dùng ngôn ngữ lập trình Matlab. Chúng tôi dùng máy tính với cấu hình sau: Chipset là Intel Xeon® CPU E5=2650 v2 @ 2.60GHz 16, hệ điều hành là Ubuntu 14.04, Ram của máy là 32G.

C. Kết quả đánh giá

Chúng tôi đã thực hiện thử nghiệm trên các bộ dữ liệu thu được để đánh giá độ chính xác của hệ thống. **Bảng 2** chỉ ra độ chính xác của hệ thống chúng tôi đề xuất cho bài toán phát hiện đồng phục. Về thời gian test trên từng ảnh, chúng tôi thực hiện lấy trung bình nhiều lần và chúng tôi nhận thấy thời gian test cho mỗi ảnh nằm trong khoảng từ 0.01 giây đến 0.04 giây. Với sự đánh giá về thời gian này, hệ thống chúng tôi đề xuất phù hợp cho bài toán phát hiện đồng phục phù hợp cho việc tải xuống các ảnh theo dòng ảnh (image stream).

Bảng 2: Độ chính xác trên các bộ dữ liệu khi thực hiện đánh giá trên hệ thống của chúng tôi

STT	Tên bộ dữ liệu	Hệ thống của chúng tôi
1	Lawson	100
2	Argentina	97.6
3	Brazil	94.0
4	Barcelona	97.0

Bên cạnh sự đánh giá về độ chính xác của hệ thống, chúng tôi cũng đánh giá với các độ đo khác để chỉ ra sự hiệu quả của hệ thống chúng tôi đề xuất cho bài toán phát hiện và phân loại đồng phục. Các độ đo khác có thể kể đến là Precision, Recall và F_1. Precision đánh giá tỉ lệ suy luận đúng của hệ thống, được tính bằng tỉ số giữa số mẫu đồng phục đoán đúng trên tổng số mẫu đồng phục được đoán.

$$precision = \frac{\{uniform\,samples\} \cap \{predicted\,uniform\,samples\}}{\{predicted\,uniform\,samples\}}$$

Recall được tính bằng tỉ số giữa số đồng phục đoán đúng trên tổng số mẫu đồng phục đang có.

$$recall = \frac{\{uniform\,samples\} \cap \{predicted\,uniform\,samples\}}{\{uniform\,samples\}}$$

F_1 được tính từ Precision và Recall.

$$F_1 = 2 * \frac{precision * recall}{precision + recall}$$

Bảng 3 và đã chỉ ra sự tin cậy của hệ thống, các độ đo nhằm khẳng định hệ thống khó có thể bỏ sót các mẫu là đồng phục, việc bỏ xót các mẫu đồng phục có thể sẽ gây ra hậu quả lớn cho công ty cũng như các doanh nghiệp.

Bảng 3: Độ chính xác trên các độ đo khi thực hiện đánh giá trên hệ thống

STT	Tên bộ dữ liệu	Precision (%)	Recall (%)	F_1 (%)
1	Lawson	100	100	100
2	Argentina	96.07	98.00	97.03
3	Brazil	92.93	92.00	92.46
4	Barcelona	96.50	95.17	95.83

5. KẾT LUẬN

Chúng tôi đã đề xuất một hệ thống hiệu quả để giải quyết bài toán nhận phát hiện đồng phục,chúng tôi đã sử dụng các phương pháp học máy, mã hóa thưa mới nhất. Đặc biệt chúng tôi cũng tập trung vào chuẩn tối ưu l_2 với mục đích để tối ưu hóa thời gian học cũng như thời gian test, trong khi vẫn đảm bảo độ chính xác cho hệ thống. Bài toán phát hiện đồng phục có thể ứng dụng trong giám sát nhân viên, quản lý hình ảnh của công ty, cũng có ý nghĩa về bảo mật thông tin cho công ty, tổ chức. Ngoài ra, trong tương lai chúng tôi sẽ áp dụng hệ thống cho việc gợi ý quảng cáo và tự động gán thẻ tag cho ảnh để phục vụ nhiều hơn cho các nhà quảng cáo, các công ty, các nhà cung cấp dịch vụ.

6. TÀI LIỆU THAM KHẢO

[1] L. Chen and A. Roy. Event detection from flickr data through waveletbased spatial analysis. In Proceedings of the 18th ACM conference on Information and knowledge management, pages 523–532. ACM, 2009.

[2] D. J. Crandall, L. Backstrom, D. Huttenlocher, and J. Kleinberg. Mapping the world's photos. In Proceedings of the 18th international conference on World wide web, pages 761–770. ACM, 2009.

[3] J. Fan, Y. Gao, and H. Luo. Multi-level annotation of natural scenes using dominant image components and semantic concepts. In Proceedings of the 12th annual ACM international conference on Multimedia, pages 540–547. ACM, 2004.

[4] S. Gu, L. Zhang, W. Zuo, and X. Feng. Projective dictionary pair learning for pattern classification. In Advances in Neural Information Processing Systems, pages 793–801, 2014.

[5] Z. Jiang, Z. Lin, and L. S. Davis. Label consistent k-svd: learning a discriminative dictionary for recognition. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 35(11):2651–2664, 2013.

[6] J. Luo, J. Yu, D. Joshi, and W. Hao. Event recognition: viewing the world with a third eye. In Proceedings of the 16th ACM international conference on Multimedia, pages 1071–1080. ACM, 2008.

[7] A. Torralba, R. Fergus, and Y. Weiss. Small codes and large image databases for recognition. In Computer Vision and Pattern Recognition, 2008. CVPR 2008. IEEE Conference on, pages 1–8. IEEE, 2008.

[8] X.-J. Wang, L. Zhang, F. Jing, and W.-Y. Ma. Annosearch: Image autoannotation by search. In Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on, volume 2, pages 1483–1490. IEEE, 2006.

[9] M. Yang, D. Zhang, and X. Feng. Fisher discrimination dictionary learning for sparse representation. In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 543–550. IEEE, 2011.

[10] Q. Zhang and B. Li. Discriminative k-svd for dictionary learning in face recognition. In Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on, pages 2691–2698. IEEE, 2010.

[11] Z. Liu. A survey on social image mining. In Intelligent Computing and Information Science, pages 662–667. Springer, 2011.

[12] P. Wu, S. C.-H. Hoi, P. Zhao, and Y. He. Mining social images with distance metric learning for automated image tagging. In Proceedings of the fourth ACM international conference on Web search and data mining, pages 197–206. ACM, 2011.

[13] J. Yu, X. Jin, J. Han, and J. Luo. Mining personal image collection for social group suggestion. In Data Mining Workshops, 2009. ICDMW'09. IEEE International Conference on, pages 202–207. IEEE, 2009.

[14] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. International journal of computer vision, 42(3):145–175, 2001.

[15] J. Wu and J. M. Rehg. Centrist: A visual descriptor for scene categorization. Pattern Analysis and Machine Intelligence, IEEE Transactions on, 33(8):1489–1501, 2011.

[16] D. G. Lowe. Object recognition from local scale-invariant features. In Computer vision, 1999. The proceedings of the seventh IEEE international conference on, volume 2, pages 1150–1157. Ieee, 1999.

[17] A. Bosch, A. Zisserman, and X. Muoz. Image classification using random forests and ferns. In Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on, pages 1–8. IEEE, 2007.

[18] H. Bay, T. Tuytelaars, and L. Van Gool. Surf: Speeded up robust features. In Computer vision–ECCV 2006, pages 404–417. Springer, 2006.

[19] M. Aharon, M. Elad, and A. Bruckstein. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. IEEE TRANSACTIONS ON SIGNAL PROCESSING, 54(11):4311, 2006.

[20] D. Zhang, M. Yang, and X. Feng. Sparse representation or collaborative representation: Which helps face recognition? In Computer Vision (ICCV), 2011 IEEE International Conference on, pages 471–478. IEEE, 2011.