KẾT HỢP ĐẶC TRƯNG CỤC BỘ VÀ NHẬN DẠNG CHỮ QUANG HỌC TRONG BÀI TOÁN TĂNG CƯỜNG HÌNH ẢNH CHO TÀI LIỆU VĂN BẢN

NGÔ QUỐC VIỆT*

TÓM TẮT

Trong bài viết này, chúng tôi trình bày kĩ thuật tăng cường mô hình hình ảnh nhằm bổ sung thêm ngữ nghĩa của các thuật ngữ hay hình ảnh phức tạp trong các tài liệu văn bản. Kĩ thuật dựa trên sự kết hợp nhận dạng chữ quang học và đặc trưng cục bộ nhằm nhận dạng các thuật ngữ, và hình vẽ có trong tài liệu văn bản để có thể giải quyết bài toán theo vết trong lĩnh vực thực tế tăng cường. Chúng tôi đã kết hợp việc khai thác và tham số hóa đặc trưng cục bộ cải tiến trên thiết bị di động để giải quyết bài toán theo vết hình ảnh thời gian thực, và kĩ thuật nhận dạng chữ quang học cho bước nhận dạng đối tượng để tăng cường mô hình thích hợp.

Từ khóa: thực tế tăng cường, nhận dạng chữ quang học, theo vết.

ABSTRACT

The combination of internal typical features and Optical Character recognition in augmenting visual models for textual documents

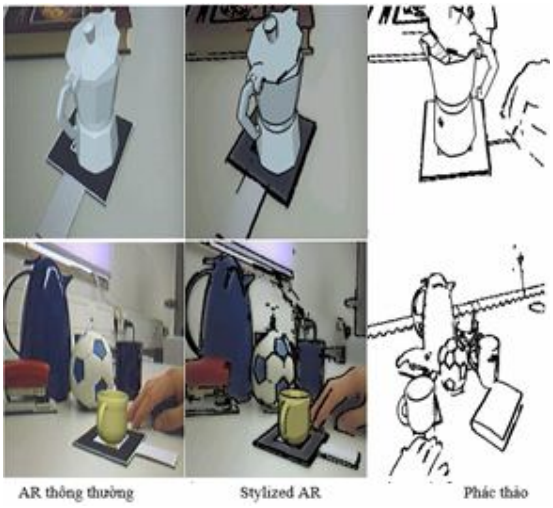
In this paper, we present the technique of augmenting the image models aiming at adding more meanings of terminologies or images in textual documents. The main technique is based on the combination of optical character recognition and internal typical features to recognize the terminologies or images in textual documents in order to solve the problem of tracking in the augmented reality field. We have incorporated the act of exploiting and digitalizing the internal typical features innovated on mobile devices so as to solve the problem of tracking images and the recognition technique of optical characters as a step for recognizing the objects and reinforcing the appropriate models.

Keywords: augmented reality, optical character recognition, ocr, tracking.

1. Giới thiệu

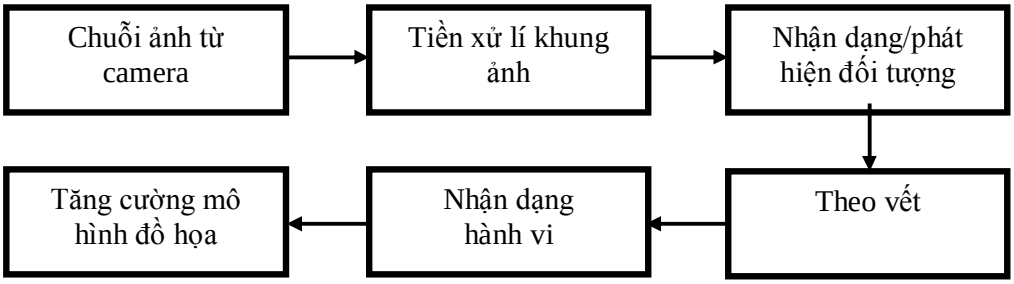
Thực tế tăng cường (**A**ugmented **R**eality) là một hướng nghiên cứu nhằm tích hợp các thông tin hay mô hình ảo vào thế giới thực làm cho người dùng có thể cảm nhận thông tin đó như được hiện hữu trong môi trường xung quanh. Thực tế tăng cường liên quan đến nhiều lĩnh vực nghiên cứu, bao gồm quá trình xử lí tín hiệu, hệ thống theo vết, đồ họa, giao diện người dùng, yếu tố con người, điện toán di động, mạng, sự hiển thị thông tin.

* TS, Trường Đại học Sư phạm TPHCM



Hình 1. *Minh họa thực tế tăng cường*

Hệ thống thực tế tăng cường cần phải đáp ứng ba điều kiện để hoạt động hợp lí. Đó là: bộ theo vết phải chính xác trong quá trình định hướng và vị trí; Bộ theo vết phải hoạt động ở nhiều môi trường (độ sáng, nhiễu, v.v..); Đáp ứng thời gian thực giữa bộ theo vết và các hình ảnh đồ họa 2D, 3D cần hiển thị trong cảnh; Một trong hai bài toán quan trọng nhất của AR là theo vết (tracking). Nhiệm vụ quan trọng của *tracking là nhận dạng đối tượng hay cảnh tự nhiên có trong các khung ảnh* từ camera, nhờ đó *tăng cường hình ảnh đồ họa thích hợp*. Ngoài ra, *cần theo vết được vị trí của đối tượng có trong khung ảnh (khi camera di chuyển, hay bản thân đối tượng di chuyển) nhằm hiển thị ảnh tăng cường ở vị trí thích hợp, hay tránh việc phải thực hiện nhận dạng cho mọi khung ảnh*. Các bước chính của AR được minh họa như sau.

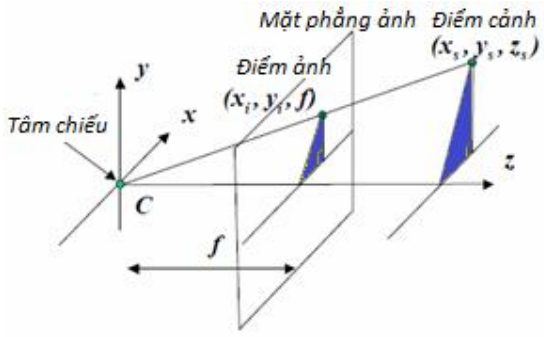


Trong bài viết này, chúng tôi trình bày kĩ thuật tăng cường mô hình hình ảnh nhằm bổ sung thêm ngữ nghĩa của các thuật ngữ hay hình ảnh phức tạp trong các tài liệu văn bản. Để thực hiện, các kĩ thuật nhận dạng đối tượng/hình ảnh dựa trên đặc trưng cục bộ được kết hợp với nhận dạng chữ quang học-OCR [1] trong bài toán theo vết nhằm xác định chính xác đối tượng hay mô hình cần tăng cường cho cảnh từ camera hay đoạn phim của thiết bị di động.

2. Xác định ma trận camera và so sánh hiệu quả của hai giải pháp theo vết dựa trên đặc trưng Fern và nhận dạng chữ quang học

2.1. Xác định tham số camera (Camera calibration)

Trong mọi trường hợp, việc xác định vị trí và hướng camera tại mỗi frame, cùng với tham số nội tại của camera nhằm render đối tượng ảo tại vị trí hợp lí trên mặt phẳng quan sát (màn hình máy quay số, màn hình máy tính, v.v..) là quan trọng. Camera calibration nhằm tìm ra vị trí và hướng của camera liên quan tới ảnh hiển thị trên màn hình quan sát. Đối tượng ảo muốn đặt hợp lí trên ảnh của màn hình quan sát, cần phải xem như được chiếu từ camera ảo (trùng với camera thật). Vấn đề này được giải quyết bằng cách sử dụng các công thức chiếu và suy luận vị trí của camera thực liên quan đến cảnh trên màn hình quan sát. Ngoài ra, việc chuẩn hóa tham số nội của camera được thực hiện nhằm xác định tình trạng biến dạng ảnh (image distortion) do môi trường hay phần cứng của camera tạo nên (thường gặp trong những camera phổ thông). Camera calibration (còn gọi là camera pose estimation) nhằm xác định ma trận 3x4 (gọi là ma trận calibration) thể hiện các tham số nội (intrinsic) và ngoại (extrinsic) của camera. Ma trận calibration thể hiện cả độ lớn tiêu cự, mức độ lệch (skew factor) ảnh, và biến dạng ống kính. Hình 2 thể hiện quan hệ giữa điểm cảnh (không gian thực) và điểm ảnh chiếu tương ứng.



Hình 2. *Pinhole camera*

Cho (x_i, y_i) là tọa độ điểm trên mặt phẳng quan sát, (x_s, y_s, z_s) là điểm 3D thế giới thực.

Ta có: $x_i = f\frac{x_s}{z_s}$, $y_i = f\frac{y_s}{z_s}, x_i = \frac{u}{w} y_i = \frac{v}{w}$, $\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix}$ (1)

Nếu tâm ảnh trên mặt phẳng chiếu không trùng với giao điểm của trục Z với mặt phẳng chiếu, thì

$$x_i = f\frac{x_s}{z_s} + t_u, y_i = f\frac{y_s}{z_s} + t_v$$

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} f & 0 & t_u \\ 0 & f & t_v \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} \quad (2)$$

Trường hợp kích thước pixel không đều, thì ma trận trên được biểu diễn bởi, với m_u, m_v là kích thước pixel theo *mm*.

$$P_c = \begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} m_u f & 0 & m_u t_u \\ 0 & m_v f & m_v t_v \\ 0 & 0 & 1 \end{pmatrix} \begin{bmatrix} x_s \\ y_s \\ z_s \end{bmatrix} = \begin{pmatrix} \alpha_x & 0 & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} P = KP \quad (3)$$

Trong một số trường hợp, cần tham số độ biến dạng ***s***, do trục *u* và*v* không trực giao. Khi đó:

$$\begin{pmatrix} u \\ v \\ w \end{pmatrix} = \begin{pmatrix} \alpha_x & s & u_0 \\ 0 & \alpha_y & v_0 \\ 0 & 0 & 1 \end{pmatrix} P = KP \quad (4)$$

K được gọi là ma trận tham số nội tại (intrinsic) của camera, với αx, αy là độ dài tiêu cự theo pixel, (u_0, v_0) tọa độ tâm ảnh theo pixel.

Nếu camera không có tâm chiếu tại (0, 0, 0) và trục *Z* không trực giao với mặt phẳng quan sát, thì cần thực hiện một phép dịch chuyển (T_x, T_y, T_z) và một phép xoay *R*. Đặt $E = (R|RT)$là tham số ngoại của camera (extrinsic parameters). Khi đó, phép biến đổi camera được xác định bởi ma trận

$$K(R \mid RT) = (KR \mid KRT) = KR(I \mid T) \quad (5)$$

P_c (điểm chiếu của *P* lên mặt phẳng quan sát) được xác định bởi

$$P_1c = KR(I \mid T)P = CP \quad (6)$$

C (kích thước3x4) được gọi là ma trận camera calibration hoàn chỉnh. Đặt $C = (KR \mid KRT) = (M \mid MT)$, với $M = KR$ (kích thước 3x3). Sử dụng phân rã RQ [2], sẽ xác định được $M = AB$, với A là ma trận tam giác trên và *B* là ma trận trực giao. Khi đó ma trận *A* ứng với *K* (tham số nội của camera), và ma trận *B* ứng với phép quay *R*. Đặt C_4 là cột cuối của ma trận C, khi đó

$$MT = C_4 \Rightarrow T = M^{-1}C_4 \quad (7)$$

Như vậy, nếu cho trước *C*, ta sẽ xác định được các tham số ngoại và nội của camera. Tuy nhiên, trong điều kiện tổng quát, các giá trị ma trận *C* không được biết trước, hoặc là những giá trị mặc định trong điều kiện nhất định, vì vậy việc xác định *C* là cần thiết.

Đặt $C = \begin{pmatrix} r_1 \\ r_2 \\ r_3 \end{pmatrix}$, với r_i là các hàng. Cho điểm 3 chiều *P* và điểm chiếu 2 chiều P_c tương ứng của P trên ảnh. Ta có:

$$u' = \frac{u}{w} = \frac{r_1.P}{r_3.P}, v' = \frac{v}{w} = \frac{r_2.P}{r_3.P} \quad (8)$$

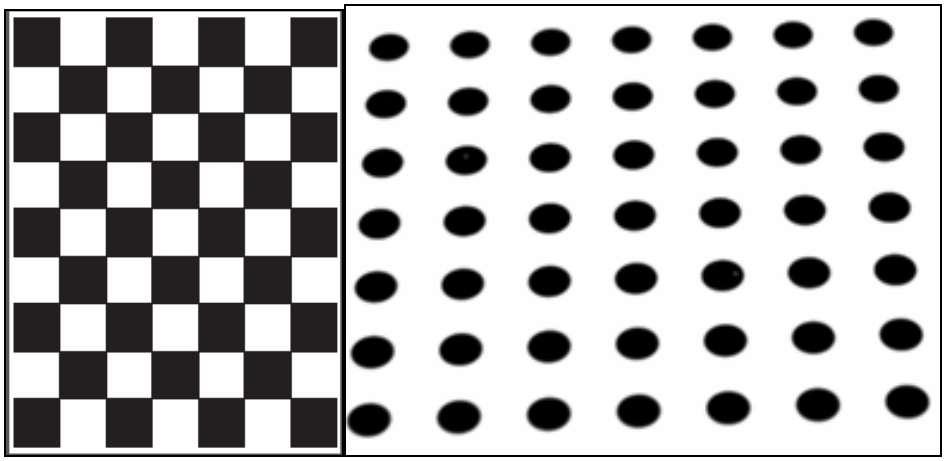
Khi đó xác định được hai phương trình

$$u'(r_3.P) - r_1.P = 0$$
$$v'(r_3.P) - r_2.P = 0 \quad (9)$$

Với mỗi cặp điểm, phát sinh được hai phương trình,vì vậy cần tối thiểu 6 cặp điểm 3D trong không gian thực và điểm 2D tương ứng trong mặt phẳng quan sát để xác định *C*.

Thủ tục camera calibration được thực hiện như sau:

B1. Tạo bản in hình grid tương tự như các mẫu sau



Hình 3. *Các mẫu dùng để xác định ma trận C*

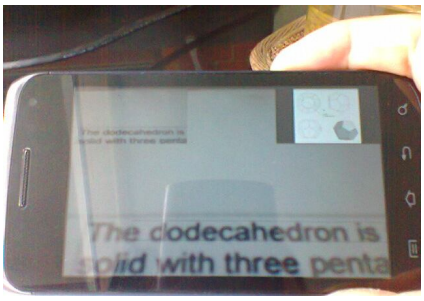
B2. Nhận dạng các điểm góc trên ảnh và điểm 3D tương ứng để giải hệ phương trình (9) nhằm tìm ra ma trận *C*.

B3. Bước B2 có thể được thực hiện nhiều lần (với camera ở các vị trí khác nhau) nhằm kiểm tra lại độ chính xác khi xác định ma trận *C*.

2.2. Thuật giải theo vết dựa trên đặc trưng Fern

Do hệ thống chỉ làm việc với các vùng (Hình 4 - chứa thuật ngữ) thay vì cả khung ảnh (chứa cả trang văn bản, có thể bao gồm hình ảnh) từ camera nên các đặc trưng (theo vết và nhận dạng) chỉ được trích trong khu vực này.

Đối với vấn đề tracking, chúng tôi thử nghiệm với hai loại đặc trưng: đặc trưng chữ quang học; và đặc trưng cục bộ Fern.



Hình 4. *Khu vực có chữ dùng cho theo vết*

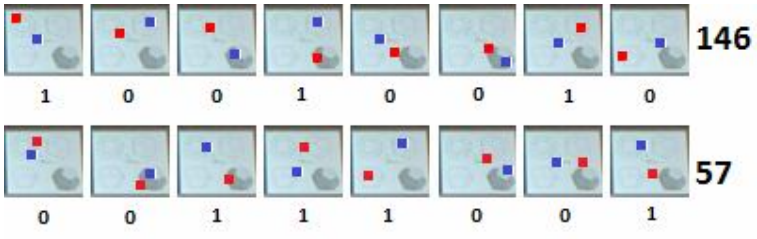
Nhằm thực hiện theo vết giữa các khung ảnh video với loại đặc trưng cục bộ, chúng tôi sử dụng hiệu chỉnh của Fern [3] gọi là PhonyFern lên toàn bộ cảnh. Phân lớp đặc trưng cho bài toán theo vết học phân bố xác suất các đặc trưng nhị phân $F(p)$ của tập các điểm mô hình m_c ứng với lớp C. Các đặc trưng nhị phân là các so sánh giữa cường độ ảnh $I(p)$ lân cận xác định bởi độ dời (l, r) so với điểm quan tâm p. $F(p)$ là 1 nếu $I(p+l) < I(p+r)$ và là 0 nếu ngược lại. Mỗi điểm quan tâm sẽ được học nhiều $F(p)$ ứng với các góc, tỉ lệ, nhiễu, ánh sáng khác nhau. N giá trị $F(p)$ được nhóm để tạo thành M *fern* có kích thước $S = N/M$ khác nhau cho mỗi điểm quan tâm. Khi đó, xác suất điều kiện của N giá trị $F(p)$ cho điểm quan tâm được xác định bởi

$$P(f_1, f_2, \ldots, f_N | C = c_i) = \prod_{k=1}^{M} P(F_k | C = c_i) \tag{10}$$

Trong thực tế, giá trị của $P(F_k|C=c_i)$ trong công thức trên có thể được xác định bởi

$$P(F_k | C = c_i) = \frac{n_{k,i} + u}{\sum_k (n_{k,i} + u)} \tag{11}$$

Với giá trị u lớn hơn 0 (được chọn bằng 1 trong thực nghiệm) nhằm làm cho giá trị *LOG* của xác suất Fern hợp lệ, và $n_{k,i}$ là giá trị *fern* được minh họa theo hình sau với kích thước *fern* là 8.



Hình 5. *Minh họa giá trị fern kích thước 8*

Khi đó xác suất của một lớp được xác định theo công thức (10) với phép nhân có thể thay thế bằng phép cộng và $P(F_k|C = c_i)$ có thể được thay thế bằng $Log\ P(F_k|C = c_i)$

Số lượng điểm quan tâm (xác định thông qua cực trị của toán tử Laplacian) có thể được giảm bớt thông qua thuật giải FAST detector [5]. Ngưỡng FAST có thể xác lập để xác định số lượng cố định các điểm quan trọng. Trong bài viết này, số điểm quan tâm (cũng là số lớp trong quá trình huấn luyện tập đặc trưng Fern) được xác lập nhỏ hơn hay bằng 320 cho ảnh kích thước 320x240. Kích thước mỗi Fern có chiều dài 8(*S=8*), số lượng Fern cho mỗi lớp là từ 32-64 (*N=256-512*). Việc sử dụng kích thước Fern bằng 8 cho phép sử dụng số nguyên 8-bit để lưu trữ giá trị xác suất cho đặc trưng Fern.

Quá trình phân lớp dựa trên các điểm quan tâm *p* được thực hiện thông qua tính toán các xác suất F_i của *p* cho mỗi Fern F_s đã lưu trữ thông qua công thức (10).

Đối với đặc trưng chữ quang học (thử nghiệm với văn bản chữ in), chúng tôi sử dụng ảnh trắng đen (đã xử lí nhiễu, góc quay) với đường viền các chữ với các đặc trưng chính được trích từ [1], [4].

Quá trình theo vết được thực nghiệm khoảng 1440 khung ảnh video cho mỗi loại ảnh, sau đó lấy thời gian trung bình cho từng loại đặc trưng OCR, Fern. Kết quả được thể hiện trong Bảng 1.

Bảng 1. *So sánh thời gian trung bình theo vết sử dụng Fern và OCR*

No	Ảnh văn bản có chứa kí tự	Fern (millisecond)	OCR (millisecond)
1	Cube	32.8	239.5
2	Tetrahedron	36.1	1080.9
3	Icosahedron	35.9	846.9
4	Octahedron	36.2	743.3
5	Sphere	38.5	174.7
6	Intersection	39.5	422.9
7	Triangular prism	40.1	490

Thời gian theo vết dựa vào nhận dạng chữ quang học là khá lớn vì đòi hỏi nhiều bước tiền xử lí như: chuyển ảnh xám; phân tích dòng văn bản, điều chỉnh dòng cơ sở; dò tìm khoảng cách từ, phân đoạn để xác định từ; nhận dạng kí tự và nhận dạng từ, trong khi đó theo vết dựa trên đặc trưng Fern có kết quả khả quan hơn nhiều. Kết quả thực nghiệm cho thấy thời gian trung bình theo vết dựa trên đặc trưng OCR không đáp ứng thời gian thực, trong khi theo vết dựa trên đặc trưng Fern đáp ứng thời gian thực cho các video frame. Độ chính xác của quá trình phát hiện và theo vết dựa vào OCR gần như tuyệt đối, trong khi độ chính xác của cùng quá trình dựa trên đặc trưng Fern phụ thuộc vào kích thước Fern và dao động từ 96% đến 99%.

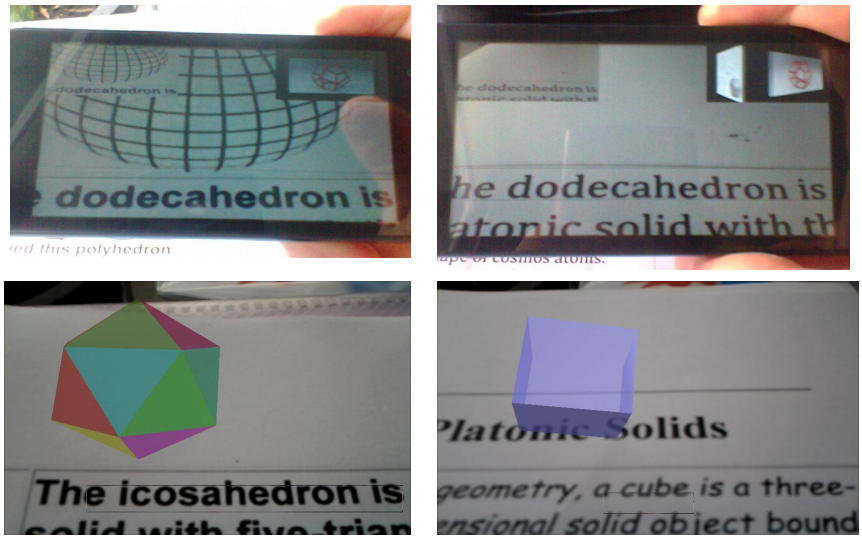
Nhận xét rằng việc kết hợp giữa đặc trưng Fern trong giải quyết bài toán theo vết kết hợp với nhận dạng chữ quang học trong bước nhận dạng chữ cho kết quả tích cực trong vấn đề tăng cường mô hình ảo cho các thuật ngữ trong các lĩnh vực khác nhau.

3. Xây dựng ứng dụng và kết quả thực nghiệm

Chúng tôi phát triển một ứng dụng minh họa trên nền Android dựa trên các thư viện mở AndAR và Tesseract OCR nhằm tạo một công cụ tương tự như một từ điển thông minh tự độngcó khả năng diễn giải các thuật ngữ thông qua các mô hình mô phỏng 3D/2D tăng cường. Các công cụ được sử dụng trong phát triển ứng dụng bao gồm:

- Bộ công cụ AndAR (*https://www.artoolworks.com/products/mobile/andar/):* hỗ trợ cho việc phát triển ứng dụng thực tế tăng cường.
- Thư viện Tesseract OCR 3.02 [6]: hỗ trợ nhận dạng chữ quang học.
- Thư viện xử lí ảnh Leptonica 1.69 (*http://www.leptonica.com/*): hỗ trợ các thao tác xử lí ảnh số.
- Thư viện OpenGL ES 1.0 (*http://www.khronos.org/opengles/*): hỗ trợ đồ họa 2D/3D trên thiết bị di động

Dữ liệu thực nghiệm được thực hiện trực tiếp trên 30 trang tài liệu văn bản in. Các trang văn bản này được trình bày bằng nhiều kiểu phông chữ khác nhau nhưng giữ nguyên kích thước chữ. Trên văn bản ngoài kí tự in còn có kèm theo các hình ảnh minh họa cho bài viết. Các ảnh tài liệu được thêm nhiễu, biến đổi *affine* với các tham số khác nhau để tạo ra tập ảnh huấn luyện cho tập đặc trưng Fern. Sau đây là một số hình kết quả của ứng dụng minh họa.



Hình 6. *Kết quả hình ảnh tăng cường cho thuật ngữ*

4. Kết luận

Trong bài viết này chúng tôi đã trình bày việc kết hợp giữa đặc trưng Fern cải tiến phù hợp cho thiết bị di động và OCR nhằm tăng cường mô hình đồ họa 2D/3D cho các thuật ngữ. Kết quả thực nghiệm với đặc trưng Fern dùng cho bước theo vết và đặc trưng vector cho bước OCR thể hiện ưu thế về tốc độ và kết quả nhận dạng chính xác so với chỉ sử dụng đặc trưng vector trên chữ. Ngoài ra, các công thức và kĩ thuật xác định ma trận camera nhằm hiển thị mô hình khớp với cảnh đã được trình bày chi tiết.

TÀI LIỆU THAM KHẢO

1. L. Eikvil (1993), "OCR - Optical Character Recognition", *Norsk regnesentral*, Norway.
2. L. El Ghaogui (2012), "Optimization Models and Applications", *UC Merkeley*.
3. M. Ozuysal, P.Fua, V. Lepetit (2007), "Fast keypoint recognition in Ten Lines of Code" CVPR'07, pp.1-8.
4. S.V.Rice, G.Nagy, T.A.Nartker (1999), "Optical Character Recognition: An Illustrated Guide to the Frontier", *Kluwer Academic Publishers*.
5. E. Rostenn, T. Drummond (2006), "Machine learning for high speed corner detection", ECCV'06, pp.430-443.
6. R. Smith (2007), "An Overview of the Tesseract OCR Engine", *Institute of Electrical and Electronics Engineers*.

(Ngày Tòa soạn nhận được bài: 29-7-2013; ngày phản biện đánh giá: 14-10-2013; ngày chấp nhận đăng: 24-10-2013)

QUÁ TRÌNH KÍCH HOẠT CỤC BỘ ĐỒNG VÀ HỢP KIM *α* -ĐỒNG THAU…

(Tiếp theo trang 168)

9. Маршаков И.К., Лесных Н.Н., Тутукина Н.М., Волкова Л.Е. (2007), "Анодное растворение меди в щелочных средах. III. Хлоридно-щелочные растворы", *Ж. конден. среды и меж. Границы*, 9(2), pp.138-141.
10. Рылкина М.В., Андреева Н. П., Кузнецов Ю. И. (1993) "Влияние pH среды на депассивацию меди", *Защита металлов*, 29(2), pp.207-222.
11. Ушакова Е.Ю., Тутукина Н.М., Маршаков И.К. (1991), "Питтинговая коррозия меди и механизм ее инициирования в карбонатно-бикарбонатных растворах", *Защита металлов*, 27(6), pp. 934-939.

(Ngày Tòa soạn nhận được bài: 01-8-2013; ngày phản biện đánh giá: 21-8-2013; ngày chấp nhận đăng: 30-8-2013)