DỊCH CỦA DEEP LEARRNING

Mục lục

Abstract 3

1. Introduction 4

2. Related Work 6

2.1. Datasets 6

2.2. Drowsiness Detection Methods 7

3. The Real-Life Drowsiness Dataset (RLDD) 8

3.1. Overview 8

3.2. Data Collection 8

3.3. Content 9

3.4. Human Judgment Baseline 10

4. The Proposed Baseline Method 10

4.1. Blink Detection and Blink Feature Extraction 11

4.2. Drowsiness Detection Pipeline 12

4.2.1. Preprocessing 12

4.2.2. Feature Transformation Layer 13

4.2.3. HM-LSTM Network 13

4.2.4. Fully Connected Layers 14

4.2.5. Regression Unit 14

4.2.6. Discretization and Voting 15

4.2.7. Loss Function 15

5. Experiments 16

5.1. Evaluation Metrics 16

5.1.1. Blink Sequence Accuracy (BSA) 16

5.1.2. Blink Sequence Regression Error (BSRE) 16

5.1.3. Video Accuracy (VA) 16

5.1.4. Video Regression Error (VRE) 17

5.2. Implementation 17

5.3. Experimental Results 17

6. Conclusions 20

7. Supplementary Material : Blink Retrieval Algorithm 21

Fig 1. 3
Fig 2. 10
Fig 3. 11
Fig 4.. 16
Fig 5. 18
Fig 6. 21
Fig 7.. 21
Fig 8. 25

A Realistic Dataset and Baseline Temporal Model for Early Drowsiness Detection

Abstract

Buồn ngủ có thể khiến tính mạng của nhiều tài xế và công nhân gặp nguy hiểm. Điều quan trọng là phải thiết kế các hệ thống thực tế và dễ triển khai trong thế giới thực để phát hiện sự khởi đầu của cơn buồn ngủ. Trong bài báo này, chúng tôi đề cập đến việc phát hiện buồn ngủ sớm, có thể cung cấp cảnh báo sớm và cung cấp cho đối tượng nhiều thời gian để phản ứng. Chúng tôi trình bày một tập dữ liệu lớn và công khai trong cuộc sống thực 1 gồm 60 đối tượng, với các phân đoạn video được gắn nhãn là cảnh giác, cảnh giác thấp hoặc buồn ngủ. Tập dữ liệu này bao gồm khoảng 30 giờ video, với nội dung từ những dấu hiệu buồn ngủ tinh vi đến những dấu hiệu rõ ràng hơn. Chúng tôi cũng chuẩn mô hình tạm thời2 cho tập dữ liệu của chúng tôi, tập dữ liệu này có nhu cầu tính toán và lưu trữ thấp. Cốt lõi của phương pháp được đề xuất của chúng tôi là mạng Bộ nhớ ngắn hạn dài hạn đa cấp độ phân cấp (HMLSTM), được cung cấp bởi các tính năng nháy mắt được phát hiện theo trình tự. Các thí nghiệm của chúng tôi chứng minh mối quan hệ giữa các đặc điểm chớp mắt liên tiếp và buồn ngủ. Trong các kết quả thử nghiệm, phương pháp cơ sở của chúng tôi tạo ra độ chính xác cao hơn so với phán đoán của con người.

Figure 1. Sample frames from the RLDD dataset in the alert (first row), low vigilant (second row) and drowsy (third row) states.

Fig 1. Các khung mẫu từ tập dữ liệu RLDD ở trạng thái thức (hàng đầu tiên), cảnh giác buồn ngủ thấp (hàng thứ hai) và buồn ngủ (hàng thứ ba).

1. Introduction

Phát hiện buồn ngủ là một vấn đề quan trọng. Các giải pháp thành công có ứng dụng trong các lĩnh vực như lái xe và nơi làm việc. Ví dụ, trong việc lái xe, Cục Quản lý An toàn Giao thông Đường cao tốc Quốc gia ở Mỹ ước tính rằng 100.000 vụ tai nạn do cảnh sát báo cáo là kết quả trực tiếp của sự mệt mỏi của người lái xe mỗi năm. Điều này dẫn đến ước tính khoảng 1.550 người chết, 71.000 người bị thương và thiệt hại tiền tệ 12,5 tỷ đô la [4]. Để thực hiện điều này, ước tính có 1 trong số 25 người lái xe trưởng thành báo cáo đã ngủ gật khi lái xe trong 30 ngày trước đó [31, 32]. Ngoài ra, các nghiên cứu chỉ ra rằng, khi lái xe trong thời gian dài, người lái xe mất khả năng tự đánh giá xem mình có buồn ngủ không [23], và đây có thể là một trong những nguyên nhân khiến nhiều vụ tai nạn xảy ra khi gần đến đích. Nghiên cứu cũng chỉ ra rằng buồn ngủ có thể ảnh hưởng đến khả năng thực hiện công việc của người lao động một cách an toàn và hiệu quả [1, 22]. Tất cả những thực tế đáng lo ngại này thúc đẩy nhu cầu về một giải pháp kinh tế có thể phát hiện tình trạng buồn ngủ trong giai đoạn đầu. Người ta thường thống nhất [29, 20, 18] rằng có ba loại nguồn thông tin trong việc phát hiện buồn ngủ: Đo hiệu suất, đo sinh lý và đo hành vi.

Ví dụ: trong lĩnh vực lái xe, các phép đo hiệu suất tập trung vào chuyển động của vô lăng, tốc độ lái, kiểu phanh và độ lệch làn đường. Một ví dụ là hệ thống Hỗ trợ chú ý của Mercedes Benz [3]. Thực tế như những phương pháp này, các công nghệ như vậy thường được dành cho các mẫu máy cao cấp, vì chúng quá đắt để người tiêu dùng bình thường có thể tiếp cận được. Các phép đo hiệu suất tại nơi làm việc có thể được thực hiện bằng cách kiểm tra thời gian phản ứng và trí nhớ ngắn hạn của công nhân [22]. Các phép đo sinh lý như nhịp tim, điện tâm đồ (ECG), điện cơ đồ (EMG), điện não đồ (EEG) [16, 28], và điện tâm đồ (EOG) [28] có thể được sử dụng để theo dõi tình trạng buồn ngủ. Tuy nhiên, các phương pháp như vậy có tính xâm nhập và không thực tế để sử dụng trong ô tô hoặc không gian làm việc mặc dù chúng có độ chính xác cao. Mũ có thể đeo được đã được đề xuất như một giải pháp thay thế cho các phép đo như vậy [2], nhưng chúng cũng không thực tế để sử dụng trong nhiều giờ.

Các phép đo hành vi thu được từ các chuyển động và biểu cảm trên khuôn mặt bằng cách sử dụng các cảm biến không xâm nhập như máy ảnh. Trong công trình của Johns [12], các thông số nháy mắt được đo bằng điốt phát quang. Tuy nhiên, phương pháp này nhạy cảm với khớp cắn, trong đó một số vật thể chẳng hạn như bàn tay được đặt giữa đi-ốt phát quang và mắt.

Máy ảnh điện thoại là một giải pháp thay thế dễ tiếp cận và rẻ cho các phương pháp nói trên. Một trong những mục tiêu của bài báo này là giới thiệu và điều tra một quy trình xử lý end-to-end sử dụng đầu vào từ camera điện thoại để phát hiện cả những dấu hiệu buồn ngủ tinh tế và được biểu hiện rõ ràng hơn trong thời gian thực. Đường ống này rẻ về mặt tính toán để cuối cùng nó có thể được triển khai như một ứng dụng điện thoại di động có sẵn cho công chúng.

Các nghiên cứu trước đây trong lĩnh vực này chủ yếu tập trung vào việc phát hiện tình trạng buồn ngủ quá mức với các dấu hiệu rõ ràng như ngáp, ngủ gật và nhắm mắt kéo dài [19, 20, 25]. Tuy nhiên, đối với

người lái xe và công nhân, những biển báo rõ ràng như vậy có thể chỉ xuất hiện cho đến khi xảy ra tai nạn. Do đó, có giá trị đáng kể trong việc phát hiện buồn ngủ ở giai đoạn đầu, để cung cấp thêm thời gian cho các phản ứng thích hợp. Bộ dữ liệu được đề xuất đại diện cho các dấu hiệu buồn ngủ tinh tế trên khuôn mặt cũng như các dấu hiệu rõ ràng và dễ quan sát hơn, do đó nó là bộ dữ liệu thích hợp để đánh giá các phương pháp phát hiện buồn ngủ sớm.

Dữ liệu của chúng tôi bao gồm khoảng 30 giờ video RGB, được ghi lại trong môi trường thực tế trong nhà bằng nhiều máy ảnh điện thoại / web khác nhau. Tốc độ khung hình dưới 30 khung hình / giây, điều này làm cho việc phát hiện buồn ngủ trở nên khó khăn hơn, vì nháy mắt không được quan sát rõ ràng như trong các video có tốc độ khung hình cao. Các video trong tập dữ liệu được gắn nhãn bằng cách sử dụng các nhãn ba lớp: cảnh giác, cảnh giác thấp và buồn ngủ (Hình 1). Các video đã được thu thập từ 60 người tham gia. Nhu cầu nghiên cứu về phát hiện buồn ngủ sớm được minh họa rõ hơn bằng các thí nghiệm mà chúng tôi đã tiến hành, trong đó chúng tôi yêu cầu hai mươi cá nhân phân loại video từ tập dữ liệu của chúng tôi thành ba lớp được xác định trước. Độ chính xác trung bình của những người quan sát là dưới 60%. Độ chính xác thấp này cho thấy bản chất khó khăn của vấn đề phát hiện buồn ngủ sớm.

Ngoài việc đóng góp một bộ dữ liệu về buồn ngủ thực tế lớn và công khai, chúng tôi cũng thực hiện một phương pháp cơ sở và bao gồm các kết quả định lượng từ phương pháp đó trong các thử nghiệm. Phương pháp được đề xuất sử dụng thông tin tạm thời của video bằng cách sử dụng mạng LSTM đa cấp độ phân cấp (HM-LSTM) [7] và biểu quyết, để lập mô hình mối quan hệ giữa chớp mắt và trạng thái tỉnh táo. Phương pháp cơ sở được đề xuất tạo ra độ chính xác cao hơn so với phán đoán của con người trong kết quả thử nghiệm của chúng tôi.

Các nghiên cứu trước đây về phát hiện buồn ngủ đã đưa ra kết quả về các tập dữ liệu là riêng tư [9] hoặc đã hoạt động [19, 20]. Bằng cách “hành động”, chúng tôi có nghĩa là dữ liệu trong đó đối tượng được hướng dẫn mô phỏng buồn ngủ, so với dữ liệu “thực tế”, chẳng hạn như dữ liệu của chúng tôi, trong đó đối tượng thực sự buồn ngủ trong các video tương ứng. Việc thiếu các bộ dữ liệu lớn, công khai và thực tế đã được các nhà nghiên cứu trong lĩnh vực này chỉ ra [18, 19, 20].

Công việc của chúng tôi được thúc đẩy ở một mức độ nào đó bởi miền lái xe (tức là góc và khoảng cách của camera trong tập dữ liệu của chúng tôi và khoảng thời gian hiệu chuẩn trong phương pháp của chúng tôi như được giải thích trong Phần 4.2). Tuy nhiên, tập dữ liệu của chúng tôi không được thu thập từ quá trình lái xe và nó không nắm bắt được một số khía cạnh quan trọng của việc lái xe như ánh sáng ban đêm và độ rung của máy ảnh do chuyển động của ô tô. Với những khía cạnh này của tập dữ liệu của chúng tôi, chúng tôi không tuyên bố rằng tập dữ liệu và kết quả của chúng tôi đại diện cho các điều kiện lái xe. Đồng thời, dữ liệu và phương pháp cơ sở được đề xuất có thể hữu ích cho các nhà nghiên cứu nhắm mục tiêu vào các ứng dụng phát hiện buồn ngủ khác, ví dụ như trong môi trường công sở.

Tập dữ liệu được đề xuất cung cấp những lợi thế đáng kể so với các tập dữ liệu công khai hiện có để phát hiện buồn ngủ, bất kể các tập dữ liệu hiện có đó có được thúc đẩy bởi miền thúc đẩy hay không: (a) đây là tập dữ liệu phát hiện buồn ngủ công khai lớn nhất cho đến nay, (b) các mẫu buồn ngủ là buồn ngủ thực sự trái ngược với buồn ngủ hành động trong [30], và (c) dữ liệu thu được bằng cách sử dụng các máy ảnh khác nhau. Mỗi đối tượng tự ghi hình bằng cách sử dụng điện thoại di động hoặc máy ảnh web của họ, trong môi trường thực tế trong nhà do họ lựa chọn. Điều này trái ngược với các bộ dữ liệu hiện có [30, 16] nơi các bản ghi được thực hiện trong bối cảnh phòng thí nghiệm, với cùng phông nền, kiểu máy ảnh và vị trí máy ảnh.

Các đóng góp khác của bài báo này có thể được tóm tắt như sau: (a) giới thiệu, như một phương pháp cơ bản, một đường dẫn phát hiện buồn ngủ theo thời gian thực đầu cuối dựa trên tốc độ khung hình thấp dẫn đến độ chính xác cao hơn so với các quan sát viên của con người, và (b) kết hợp các tính năng nháy mắt với Mạng thần kinh tái tạo đa cấp phân cấp để xử lý phát hiện buồn ngủ bằng cách sử dụng các dấu hiệu tinh tế. Những dấu hiệu này, có thể dễ dàng bị bỏ qua bởi những người quan sát của con người, rất hữu ích để phát hiện sự xuất hiện của cơn buồn ngủ ở giai đoạn đầu trước khi nó đạt đến mức nguy hiểm.

2. Related Work

Phát hiện buồn ngủ đã được nghiên cứu trong nhiều năm. Trong phần còn lại của phần này, phần đánh giá về các bộ dữ liệu có sẵn và các phương pháp hiện có sẽ được cung cấp.

2.1. Datasets

Như đã chỉ ra ở trên, có rất nhiều công trình phát hiện buồn ngủ, nhưng không có công trình nào sử dụng tập dữ liệu công khai và thực tế. Do đó, rất khó để so sánh các phương pháp trước đây với nhau và để quyết định tình trạng nghệ thuật trong lĩnh vực này là gì. Một số phương pháp hiện có [12, 17, 27, 29, 35] được đánh giá trên một số ít đối tượng mà không chia sẻ video. Trong một số trường hợp [11, 20] các đối tượng được hướng dẫn hành động buồn ngủ, trái ngược với việc lấy dữ liệu từ các đối tượng thực sự buồn ngủ.

Một số bộ dữ liệu [34, 33, 15] đã được tạo để phát hiện biểu hiện vi mô ngắn và chung chung, không áp dụng cụ thể cho việc phát hiện buồn ngủ. Tập dữ liệu phát hiện buồn ngủ NTHUdriver là tập dữ liệu công khai chứa các video IR của 36 người tham gia trong khi họ mô phỏng lái xe [30]. Tuy nhiên, nó dựa trên các đối tượng giả vờ buồn ngủ và đó là một câu hỏi mở liệu các video giả vờ buồn ngủ ở mức độ nào và ở mức độ nào có phải là dữ liệu đào tạo hữu ích để phát hiện buồn ngủ thực sự, đặc biệt là ở giai đoạn đầu hay không.

Bộ dữ liệu DROZY [16], chứa nhiều loại dữ liệu liên quan đến buồn ngủ bao gồm các tín hiệu như điện não đồ, EOG và hình ảnh hồng ngoại gần (NIR). Một lợi thế của tập dữ liệu DROZY là dữ liệu buồn ngủ được thu thập bởi các đối tượng thực sự buồn ngủ, trái ngược với việc giả vờ buồn ngủ. So với tập dữ liệu DROZY, tập dữ liệu của chúng tôi có ba ưu điểm: Thứ nhất, chúng tôi có số lượng đối tượng lớn hơn đáng kể (60 so với 14). Thứ hai, đối với mỗi đối tượng, chúng tôi có dữ liệu hiển thị đối tượng đó trong

mỗi ba lớp cảnh giác được xác định trước, trong khi trong tập dữ liệu DROZY, một số đối tượng không được ghi lại ở cả ba trạng thái. Thứ ba, trong DROZY, tất cả các video được quay bằng cách sử dụng cùng một vị trí máy ảnh và phông nền, trong điều kiện phòng thí nghiệm được kiểm soát, trong khi trong tập dữ liệu của chúng tôi, mỗi đối tượng sử dụng điện thoại di động của họ và phông nền khác nhau. So với DROZY, tập dữ liệu của chúng tôi cũng có sự khác biệt quan trọng là nó cung cấp video màu, trong khi DROZY cung cấp một số phương thức khác, nhưng chỉ video NIR.

2.2. Drowsiness Detection Methods

Cuối cùng nhưng không kém phần quan trọng, Friedrichs và Yang [9], đã sử dụng 90 giờ lái xe thực tế để đào tạo và đánh giá phương pháp của họ, nhưng tập dữ liệu của họ là riêng tư và không có sẵn để làm tiêu chuẩn.

Các tính năng phát hiện buồn ngủ không xâm nhập bằng camera được chia thành các tính năng thủ công hoặc các tính năng được học tự động bằng CNN. Về các tính năng thủ công, vùng trên khuôn mặt có nhiều thông tin nhất về tình trạng buồn ngủ là mắt và các đặc điểm thường được sử dụng thường liên quan đến hành vi chớp mắt. McIntire và cộng sự. [17] chỉ ra cách tần suất và thời lượng chớp mắt thường tăng lên khi mệt mỏi bằng cách đo thời gian phản ứng và sử dụng máy theo dõi mắt. Svensson [28] đã chỉ ra rằng biên độ chớp mắt cũng có thể là một yếu tố quan trọng. Friedrichs và Yang [9] nghiên cứu nhiều đặc điểm của chớp mắt như vận tốc mở mắt, tốc độ nhắm mắt trung bình, thời lượng chớp mắt, thời gian ngủ nhỏ và năng lượng của chớp mắt cũng như thông tin chuyển động của đầu. Họ báo cáo tỷ lệ phân loại cuối cùng là 82,5% trên tập dữ liệu riêng tư của họ, lớn hơn đáng kể so với độ chính xác 65,2% mà chúng tôi báo cáo trong các thử nghiệm của mình. Tuy nhiên, tất cả các tính năng trong [9] được trích xuất bằng cảm biến Máy nhìn thấy [5] không chỉ sử dụng thông tin video (với tốc độ khung hình 60 khung hình / giây) mà còn cả tốc độ của ô tô, thông tin GPS và tín hiệu chuyển động của đầu để phát hiện buồn ngủ. Ngược lại, trong công việc của chúng tôi, dữ liệu đến từ điện thoại di động / máy ảnh web.

Nghiên cứu gần đây kiểm tra tính hiệu quả của Mạng thần kinh sâu trong việc trích xuất tính năng đầu cuối và phát hiện buồn ngủ, trái ngược với các công trình sử dụng các tính năng thủ công với bộ phân loại hoặc hồi quy thông thường như phân tích hồi quy và phân biệt (LDA) [27], hoặc điều chỉnh một Gaussian 2D với ngưỡng [12]. Kết quả của các nghiên cứu được đề cập không được xác nhận dựa trên một tập dữ liệu lớn hoặc công khai.

Park và cộng sự. [19] tinh chỉnh ba CNN và áp dụng SVM cho các tính năng kết hợp của ba mạng đó để phân loại mỗi khung hình thành bốn loại thức, ngáp, gật đầu và buồn ngủ kèm theo nháy mắt. Mô hình được đào tạo dựa trên tập dữ liệu buồn ngủ NTHU dựa trên trạng thái buồn ngủ giả vờ và được thử nghiệm trên phần đánh giá của tập dữ liệu NTHU bao gồm 20 video chỉ của bốn người, dẫn đến độ chính xác phát hiện buồn ngủ là 73%. Chúng tôi cần lưu ý rằng độ chính xác mà chúng tôi báo cáo trong thử nghiệm của mình là 65,2%, thấp hơn độ chính xác 73% được báo cáo trong [19]. Tuy nhiên, phương pháp [19] được đánh giá dựa trên dữ liệu giả, trong đó các dấu hiệu buồn ngủ có xu hướng dễ dàng nhìn thấy và thậm chí phóng đại. Ngoài ra, công trình của Park et al. không xem xét việc tổng hợp thông tin

thời gian trong video và phân loại từng khung hình một cách độc lập, do đó nó chỉ có thể phân loại dựa trên các dấu hiệu buồn ngủ rõ ràng.

Bhargava và cộng sự. [20] chỉ ra cách một mạng lưới sâu được chắt lọc có thể được sử dụng cho các hệ thống nhúng. Điều này liên quan đến phương pháp cơ sở được đề xuất trong bài báo này, phương pháp này cũng nhằm mục đích cho các yêu cầu tính toán thấp. Độ chính xác được báo cáo trong [20] là 89% khi sử dụng ba lớp (cảnh giác, ngáp, buồn ngủ), dựa trên đào tạo về các mảng mắt và môi. Tương tự như công việc của Park và cộng sự, mạng của Bhargava và cộng sự cũng phân loại từng khung một cách độc lập, do đó không sử dụng các tính năng tạm thời. Tập dữ liệu mà họ sử dụng là riêng tư và dựa trên tình trạng ngủ gật, vì vậy rất khó để so sánh những kết quả đó với kết quả được báo cáo trong bài báo này.

3. The Real-Life Drowsiness Dataset (RLDD)

3.1. Overview

Tập dữ liệu RLDD được tạo ra cho nhiệm vụ phát hiện buồn ngủ nhiều tầng, nhắm mục tiêu không chỉ các trường hợp cực đoan và dễ nhìn thấy mà còn cả các trường hợp buồn ngủ tinh vi. Việc phát hiện những trường hợp tinh vi này có thể rất quan trọng để phát hiện buồn ngủ ở giai đoạn đầu, để kích hoạt các cơ chế ngăn ngừa buồn ngủ. Tập dữ liệu RLDD của chúng tôi là tập dữ liệu về tình trạng buồn ngủ thực tế lớn nhất cho đến nay.

Tập dữ liệu RLDD bao gồm khoảng 30 giờ video RGB của 60 người tham gia khỏe mạnh. Đối với mỗi người tham gia, chúng tôi thu được một video cho mỗi ba lớp khác nhau: tỉnh táo, cảnh giác kém và buồn ngủ, trong tổng số 180 video. Đối tượng là sinh viên đại học hoặc sau đại học và nhân viên tham gia tự nguyện hoặc khi nhận thêm tín chỉ trong một khóa học. Tất cả những người tham gia đều trên 18 tuổi. Có 51 đàn ông và 9 phụ nữ, thuộc các sắc tộc khác nhau (10 người da trắng, 5 người gốc Tây Ban Nha không phải da trắng, 30 người Indo-Aryan và Dravidian, 8 người Trung Đông và 7 người Đông Á) và độ tuổi (từ 20 đến 59 tuổi với trung bình là 25 và độ lệch chuẩn là 6). Các đối tượng đeo kính ở 21 trong số 180 video và có lông mặt đáng kể ở 72 trong số 180 video. Các video được quay từ nhiều góc độ khác nhau trong các môi trường và bối cảnh thực tế khác nhau. Mỗi video do người tham gia tự quay,

sử dụng điện thoại di động hoặc máy ảnh web của họ. Tốc độ khung hình luôn nhỏ hơn 30 khung hình / giây, đại diện cho tốc độ khung hình mong đợi của các máy ảnh thông thường được dân số chung sử dụng.

3.2. Data Collection

Trong phần này, chúng tôi mô tả cách chúng tôi thu thập các video cho tập dữ liệu RLDD. 60 người khỏe mạnh đã tham gia thu thập dữ liệu. Sau khi ký vào mẫu chấp thuận, các đối tượng được hướng dẫn quay ba video về chính họ bằng điện thoại / máy ảnh web (thuộc bất kỳ kiểu máy hoặc loại nào) ở ba trạng thái buồn ngủ khác nhau, dựa trên bảng KSS [6] (Bảng 1), trong khoảng 10 mỗi phút. Các đối tượng được yêu cầu tải lên các video cũng như nhãn tương ứng của họ trên một cổng thông tin trực tuyến được cung cấp thông qua một liên kết. Các đối tượng đã có nhiều thời gian (20 ngày) để sản xuất ba video. Hơn nữa, họ được tự do quay video ở nhà hoặc ở trường đại học, bất cứ lúc nào họ cảm thấy tỉnh táo, kém cảnh giác hoặc buồn ngủ trong khi vẫn thiết lập máy ảnh (góc và khoảng cách) gần giống nhau. Tất

cả các video đều được quay ở góc sao cho cả hai mắt đều có thể nhìn thấy được và máy ảnh được đặt trong khoảng cách một sải tay từ đối tượng. Các hướng dẫn này được sử dụng để tạo các video tương tự như video thu được trên ô tô, bằng điện thoại được đặt trong giá đỡ điện thoại trên bảng điều khiển của ô tô khi lái xe. Thiết lập được đề xuất là đặt điện thoại đối diện với màn hình máy tính xách tay của họ khi họ đang xem hoặc đọc nội dung gì đó trên máy tính của họ. Sau khi một người tham gia tải ba video lên, chúng tôi đã xem toàn bộ video để xác minh tính xác thực của chúng và để đảm bảo rằng các hướng dẫn của chúng tôi đã được tuân theo. Trong trường hợp có bất kỳ câu hỏi nào, chúng tôi đã liên hệ với những người tham gia và yêu cầu họ chia sẻ thêm chi tiết về tình huống mà họ đã quay từng video. Trong một số trường hợp, chúng tôi đã yêu cầu họ làm lại bản ghi và nếu video rõ ràng không thực tế (mọi người giả vờ buồn ngủ thay vì buồn ngủ) hoặc tiêu chuẩn, chúng tôi chỉ bỏ qua những video đó vì lý do chất lượng. Ba lớp học đã được giải thích cho những người tham gia như sau:

1. **Thức**: Một trong ba trạng thái đầu tiên được đánh dấu trong bảng KSS ở Bảng 1. Đối tượng được cho biết rằng tỉnh táo có nghĩa là họ không có dấu hiệu buồn ngủ.
2. **Cảnh giác thấp**: Như đã nêu ở mức 6 và 7 của Bảng 1, trạng thái này tương ứng với các trường hợp tinh tế khi một số dấu hiệu buồn ngủ xuất hiện, hoặc buồn ngủ nhưng không cần cố gắng giữ tỉnh táo.
3. **Buồn ngủ**: Trạng thái này có nghĩa là đối tượng cần cố gắng tích cực để không buồn ngủ (mức 8 và 9 trong Bảng 1).

1- Extremely alert
2- Very alert
3- Alert
4- Rather alert
5- Neither alert nor sleepy
6- Some signs of sleepiness
7- Sleepy, no difficulty remaining awake
8- Sleepy, some effort to keep alert
9- Extremely sleepy, fighting sleep

Table 1. KSS drowsiness scale

3.3. Content

Tập dữ liệu này bao gồm 180 video RGB. Mỗi video dài khoảng 10 phút và được gắn nhãn là thuộc một trong ba loại: thức (được gắn nhãn là 0), cảnh giác thấp (được dán nhãn là 5) và buồn ngủ (được dán nhãn là 10). Các nhãn do chính những người tham gia cung cấp, dựa trên trạng thái chủ yếu của họ trong khi quay mỗi video. Rõ ràng, có một yếu tố chủ quan trong việc quyết định các nhãn này, nhưng chúng tôi đã không tìm ra cách tốt để khắc phục vấn đề đó, do không có bất kỳ cảm biến nào có thể cung cấp một thước đo khách quan về mức độ thức. Loại nhãn này có tính đến và nhấn mạnh sự chuyển đổi từ

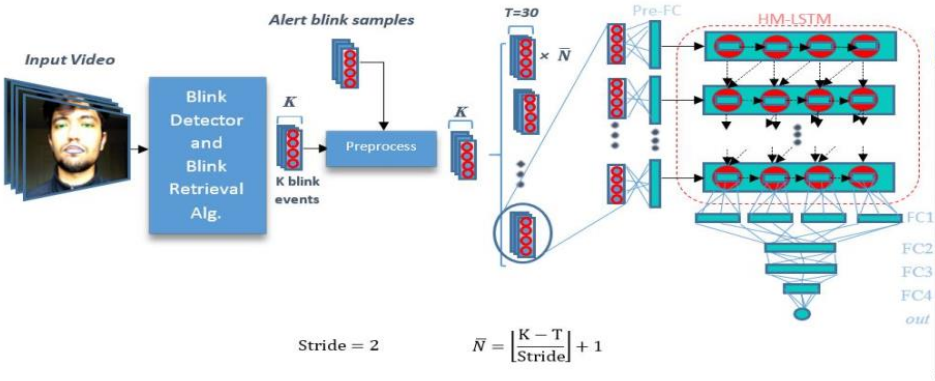
trạng thái tỉnh táo sang buồn ngủ. Mỗi bộ video được quay bằng điện thoại di động cá nhân hoặc máy ảnh web dẫn đến chất lượng và độ phân giải video khác nhau. 60 đối tượng được chia ngẫu nhiên thành năm lần gồm 12 người tham gia, nhằm mục đích xác nhận chéo. Tập dữ liệu có tổng kích thước là 111,3 Gigabyte.

3.4. Human Judgment Baseline

Chúng tôi đã tiến hành một bộ thí nghiệm để đo lường khả năng phán đoán của con người trong việc phát hiện buồn ngủ nhiều giai đoạn. Trong các thử nghiệm này, chúng tôi đã yêu cầu bốn tình nguyện viên mỗi màn hình (tổng cộng 20 tình nguyện viên) xem các video không gắn nhãn và tắt tiếng trong mỗi màn hình và viết ra một số thực từ 0 đến 10 ước tính mức độ buồn ngủ trên mỗi video (xem Bảng 1). Trước khi thử nghiệm, các tình nguyện viên (8 nữ và 12 nam, 3 sinh viên đại học và 17 nghiên cứu sinh) đã được xem một số đoạn video mẫu minh họa thang điểm buồn ngủ. Sau đó, họ được để một mình trong phòng để xem video (họ được phép tua lại hoặc tua đi các video tùy ý) và chú thích chúng. Để đảm bảo rằng mỗi nhận định độc lập với các video khác của cùng một người, các tình nguyện viên được hướng dẫn chú thích một video của mỗi chủ đề trước khi chú thích video thứ hai cho bất kỳ chủ đề nào. Kết quả của các thí nghiệm này được trình bày trong phần 5.3 và so sánh với kết quả của phương pháp cơ sở của chúng tôi. Các quan sát viên (26,1 ± 2,9 tuổi (trung bình ± SD)) đến từ các chuyên ngành khoa học máy tính, tâm lý học, điều dưỡng, công tác xã hội và hệ thống thông tin.

4. The Proposed Baseline Method

Trong phần này, chúng tôi thảo luận về các thành phần riêng lẻ của quy trình phát hiện buồn ngủ nhiều giai đoạn được đề xuất của chúng tôi. Tính năng phát hiện nháy mắt và trích xuất tính năng nháy mắt được mô tả đầu tiên. Sau đó, chúng tôi thảo luận về cách chúng tôi tích hợp mô-đun LSTM Đa cấp phân cấp vào mô hình của mình, cách chúng tôi hình thành tính năng phát hiện buồn ngủ ban đầu như một vấn đề hồi quy và cách chúng tôi tùy chỉnh đầu ra hồi quy để có được nhãn phân loại cho mỗi đoạn video. Cuối cùng, chúng tôi thảo luận về quy trình bỏ phiếu được áp dụng trên kết quả phân loại của tất cả các phân đoạn của video.

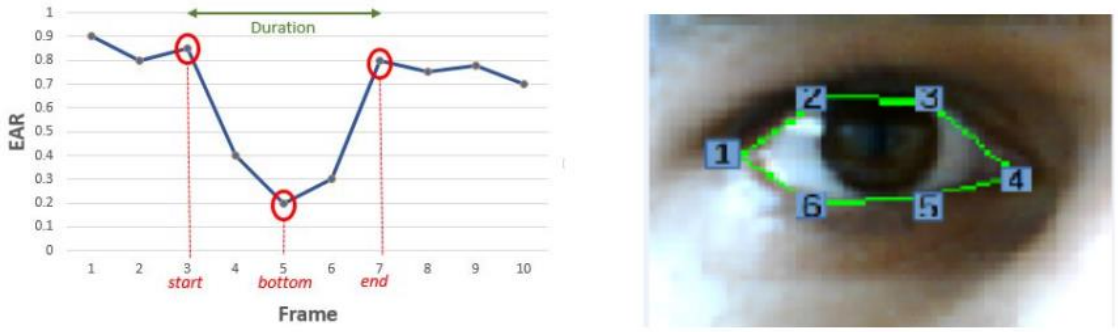


Layer	#Units	#Parameters
Pre-FC	L = 32	4x32+32
HMLSTM	$H = 32$	3x (97x(32x4+1))+ (65x(32x4+1))
FC1	$L_1 = 16$	4x(32x16+16)
FC2	$L_2 = 16$	4x16x16+16
FC3	$L_3 = 16$	16x16+16
FC4	$L_4 = 8$	16x8+8
out	1	8x1+1

Figure 2. The model design and configuration.

Fig 2. Thiết kế và cấu hình mô hình

4.1. Blink Detection and Blink Feature Extraction



(a) (b)

Figure 3. (a) The EAR sequence during an entire blink and the start, bottom and end points. (b) The eye landmarks to define EAR for each frame.

Fig 3. (a) Chuỗi EAR trong toàn bộ thời gian nháy mắt và các điểm bắt đầu, điểm cuối và điểm kết thúc. (b) Các mốc mắt để xác định EAR cho mỗi khung hình

Động lực đằng sau việc sử dụng các tính năng liên quan đến chớp mắt như thời lượng chớp mắt, biên độ và vận tốc mở mắt, là để nắm bắt các mô hình thời gian xuất hiện tự nhiên trong mắt người và có thể bị các công cụ dò tìm đặc điểm không gian như CNNs bỏ qua (như trường hợp của con người tầm nhìn được thể hiện trong các thí nghiệm của chúng tôi). Chúng tôi đã sử dụng máy dò khuôn mặt được đào tạo trước của dlib dựa trên một sửa đổi đối với Biểu đồ tiêu chuẩn của Định hướng Gradients + phương pháp SVM tuyến tính để phát hiện đối tượng [8].

Chúng tôi đã cải tiến thuật toán của Soukupova và Cech [´26] để phát hiện nháy mắt, sử dụng sáu điểm mốc trên khuôn mặt cho mỗi mắt được mô tả trong [13] để trích xuất các nháy nhanh liên tiếp mà ban đầu bị bỏ qua trong công việc của Soukupova và Cech. Máy dò mốc khuôn mặt [13] của Kazemi và Sullivan được đào tạo dựa trên "bộ dữ liệu trong tự nhiên", do đó nó hoạt động mạnh mẽ hơn với các mức độ chiếu sáng khác nhau, các biểu cảm khuôn mặt khác nhau và xoay đầu không phải trực diện vừa phải, so với việc khớp tương quan với các mẫu mắt hoặc phép chiếu cường độ hình ảnh theo chiều ngang hoặc chiều dọc theo phương pháp heuristic [26]. Trong các thí nghiệm của mình, chúng tôi nhận thấy rằng phương pháp [26] thường phát hiện các lần nháy mắt liên tiếp chỉ là một lần chớp mắt. Điều này đã tạo ra một vấn đề cho các bước phát hiện buồn ngủ tiếp theo vì nhiều lần chớp mắt liên tiếp có thể là dấu hiệu của buồn ngủ. Chúng tôi đã thêm một bước xử lý sau (Thuật toán thu hồi nháy mắt) và áp dụng nó ở phía trên đầu ra của [26], để xác định thành công nhiều lần nháy mắt có thể xuất hiện trong một lần phát hiện do [26] tạo ra. Bước xử lý hậu kỳ của chúng tôi, mặc dù dài dòng để mô tả, nhưng lại dựa vào phương pháp phỏng đoán và không cấu thành một đóng góp nghiên cứu. Để cho phép các kết quả của chúng tôi được sao chép, chúng tôi cung cấp các chi tiết của bước xử lý sau đó làm tài liệu bổ sung.

Đầu vào cho mô-đun phát hiện nháy mắt là toàn bộ video (với độ dài khoảng mười phút trong tập dữ liệu của chúng tôi). Trong một ứng dụng phát hiện buồn ngủ trong thế giới thực, nơi bạn nên đưa ra quyết định sau mỗi vài phút, đầu vào có thể chỉ bao gồm vài phút cuối cùng của video. Đầu ra của mô-đun phát hiện nháy mắt là một chuỗi các sự kiện nháy mắt {$\mathbf{blink}_1$, …, $\mathbf{blink}_k$}. Mỗi $\mathbf{blink}_i$ là một vectơ bốn chiều chứa bốn đặc điểm mô tả nháy mắt: thời lượng, biên độ, vận tốc mở mắt và tần số. Đối với mỗi sự kiện $\mathbf{blink}_i$, chúng tôi xác định $start_i$, $bottom_i$ và end_i là các điểm "bắt đầu", "đáy" và "kết thúc" (khung) trong nháy mắt đó (Hình.3a) được giải thích trong Thuật toán thu hồi nháy mắt. Ngoài ra, đối với mỗi khung *k*, chúng tôi ký hiệu:

$$EAR[k] = \frac{||\vec{p_2} - \vec{p_6}|| + ||\vec{p_3} - \vec{p_5}||}{||\vec{p_1} - \vec{p_4}||} \quad (1)$$

trong đó $\vec{p_t}$ là vị trí 2D của điểm mốc trên khuôn mặt tính từ vùng mắt (Hình 3b). Sử dụng ký hiệu này, chúng tôi xác định bốn đặc trưng bất biến quy mô chính mà chúng tôi trích xuất từ $\mathbf{blink}_i$. Đây là những tính năng mà chúng tôi sử dụng cho phương pháp phát hiện buồn ngủ ban đầu của mình

$$\text{Duration}_i = end_i - start_i + 1 \quad (2)$$

$$\text{Amplitude}_i = \frac{EAR[start_i] - 2EAR[bottom_i] + EAR[end_i]}{2} \quad (3)$$

$$\text{Eye Opening Velocity}_i = \frac{EAR[end_i] - EAR[bottom_i]}{end_i - bottom_i} \quad (4)$$

$$\text{Frequency}_i = 100 \times \frac{\text{Number of blinks up to blink}_i}{\text{Number of frames up to } end_i} \quad (5)$$

4.2. Drowsiness Detection Pipeline

4.2.1. Preprocessing

Một thách thức lớn trong việc sử dụng các tính năng chớp mắt để phát hiện buồn ngủ là sự khác biệt về kiểu chớp mắt giữa các cá nhân [9, 11, 28, 21], do đó, các tính năng nên được chuẩn hóa trên các đối tượng nếu chúng ta định huấn luyện toàn bộ dữ liệu cùng một lúc. Để giải quyết thách thức này, chúng tôi sử dụng một phần ba số lần nháy mắt đầu tiên của **trạng thái thức** để tính giá trị trung bình và độ lệch chuẩn của từng tính năng cho mỗi cá nhân, sau đó sử dụng Công thức 6 để chuẩn hóa phần còn lại của trạng thái thức nháy mắt cũng như nháy mắt ở hai trạng thái khác của cùng một người (m) và đặc trưng (n):

$$\overline{\text{Feature}}_{n,m} = \frac{\text{Feature}_{n,m} - \mu_{n,m}}{\sigma_{n,m}} \qquad (6)$$

Ở đây, $\boldsymbol{\mu_{n,m}}$ và $\boldsymbol{\sigma_{n,m}}$ là giá trị trung bình và độ lệch chuẩn của đặc trưng n trong một phần ba số lần nháy mắt đầu tiên của video trạng thái thức đối với đối tượng m.

Chúng tôi thực hiện việc chuẩn hóa này cho cả dữ liệu đào tạo và kiểm tra của tất cả các môn học và tính năng. Một cách tiếp cận tương tự đã được thực hiện trong [11, 28]. Việc chuẩn hóa này là một hạn chế thực tế: khi người lái xe bắt đầu lái một chiếc ô tô mới hoặc một công nhân bắt đầu làm việc, máy ảnh có thể sử dụng vài phút đầu tiên (trong thời gian người đó dự kiến sẽ cảnh giác) để tính toán giá trị trung bình và phương sai, đồng thời hiệu chỉnh hệ thống. Hiệu chuẩn này có thể được sử dụng cho tất cả các chuyến đi hoặc phiên tiếp theo. Máy dò quyết định trạng thái của đối tượng liên quan đến số liệu thống kê được thu thập trong giai đoạn hiệu chuẩn. Chúng tôi nên làm rõ rằng, trong các thử nghiệm của chúng tôi, các nháy mắt trạng thái thức được sử dụng để chuẩn hóa sẽ không bao giờ được sử dụng lại cho đào tạo hoặc kiểm tra. Sau quá trình chuẩn hóa cho từng cá nhân, chúng tôi thực hiện bước chuẩn hóa thứ hai, trong đó chúng tôi chuẩn hóa từng đối tượng địa lý để phân bố đối tượng địa lý có giá trị trung bình bằng 0 và phương sai là một.

4.2.2. Feature Transformation Layer

Thay vì xác định ban đầu một số lượng lớn các đối tượng địa lý, và sau đó chọn những đối tượng phù hợp nhất [9], chúng tôi để mạng sử dụng bốn đối tượng địa lý nháy mắt chính và học cách ánh xạ chúng tới không gian đối tượng có chiều cao hơn để giảm thiểu hàm mất mát. Mục tiêu của lớp được kết nối đầy đủ trước mô-đun HM-LSTM là lấy từng vectơ đặc trưng 4D tại mỗi bước thời gian làm đầu vào và biến đổi nó thành không gian chiều L với trọng số được chia sẻ ($W \in \mathbb{R}^{4\times L}$) và biases ($b \in \mathbb{R}^{1\times L}$) qua các bước thời gian. Xác định T là số bước thời gian được sử dụng cho Mạng HM-LSTM và $f_i \in \mathbb{R}^{1\times L}$ cho mỗi lần nháy mắt tại mỗi bước thời gian i, sao cho:

$$F = \text{ReLU}(BW + \bar{b}) \qquad (7)$$

trong đó $\mathbf{F} = [\mathbf{f_1}^T, \mathbf{f_2}^T, \ldots, \mathbf{f_T}^T]^T$;

$\bar{\boldsymbol{b}} = [\mathbf{b}^T, \mathbf{b}^T, \ldots, \mathbf{b}^T]^T$, $\bar{\boldsymbol{b}} \in \mathbb{R}^{T\times l}$

và $\mathbf{B} = [\mathbf{blink_1}^T, \mathbf{blink_1}^T, \ldots, \mathbf{blink_T}^T]^T$

4.2.3. HM-LSTM Network

Cách tiếp cận của chúng tôi giới thiệu một mô hình thời gian để phát hiện cơn buồn ngủ. Công trình của [29], sử dụng Mô hình Markov ẩn (HMM), gợi ý rằng các đặc điểm buồn ngủ tuân theo một mô hình

theo thời gian. Do đó, chúng tôi đã sử dụng mạng HMLSTM [7] để tận dụng mô hình thời gian trong nháy mắt. Người ta cũng mơ hồ rằng mỗi lần chớp mắt có liên quan đến các lần chớp mắt khác như thế nào hoặc bao nhiêu lần chớp mắt liên tiếp có thể ảnh hưởng lẫn nhau. Để khắc phục thách thức này, chúng tôi đã sử dụng các tế bào HM-LSTM để khám phá cấu trúc phân cấp cơ bản trong một trình tự chớp mắt.

Chung và cộng sự. [7] giới thiệu một bộ dò biên được tham số hóa, cho ra giá trị nhị phân, trong mỗi lớp của RNN xếp chồng lên nhau. Đối với bộ dò biên này, đầu ra tích cực cho một lớp tại một bước thời gian cụ thể biểu thị sự kết thúc của một phân đoạn tương ứng với mức trừu tượng tiềm ẩn cho lớp đó. Mỗi trạng thái ô được "cập nhật", "sao chép" hoặc "xóa" dựa trên các giá trị của bộ dò ranh giới liền kề. Do đó, mạng HM-LSTM có xu hướng tìm hiểu thang thời gian tốt cho các lớp cấp thấp và thang thời gian thô cho các lớp cấp cao. Phân tích thứ bậc động này cho phép mạng xem xét các lần nháy mắt cả trong các phân đoạn ngắn và dài, tùy thuộc vào thời điểm bộ dò biên được kích hoạt cho mỗi ô. Để biết thêm chi tiết về HM-LSTM, chúng tôi giới thiệu độc giả đến [7].

Mạng HM-LSTM lấy mỗi hàng F làm đầu vào tại mỗi bước thời gian và xuất ra trạng thái ẩn $h_l \in \mathbb{R}^{1\times H}$ chỉ ở bước thời gian cuối cùng cho mỗi lớp l. H là số trạng thái ẩn trên mỗi lớp.

4.2.4. Fully Connected Layers

Chúng tôi đã thêm một lớp được kết nối đầy đủ (với $W_{1,l} \in \mathbb{R}^{H\times L1}$ làm trọng số và $b_{1,l} \in \mathbb{R}^{1\times L1}$ làm biases) vào đầu ra của mỗi lớp l với các đơn vị L1 để thu thập kết quả của mạng HM-LSTM từ các các quan điểm thứ bậc một cách riêng biệt. Xác định $e_{1l} \in \mathbb{R}^{1\times L1}$ cho mỗi lớp, sao cho:

$$\mathbf{e}_{1l} = \mathrm{ReLU}(\mathbf{h}_l W_{1,l} + \mathbf{b}_{1,l}) \qquad (8)$$

Sau đó, ta ghép e_{1l} với l $\in \{1, 2, \dots, \bar{L}\}$ để tạo thành e1 = [$\mathbf{e_{11}}$, $\mathbf{e_{12}}$, …, $\boldsymbol{e_{1\bar{L}}}$], trong đó $e_1 \in \mathbb{R}^{1\times(L_1\cdot\bar{L})}$ và $\bar{L}$ là số lớp.

Tương tự, như trong Hình 2, e1 được cấp cho các lớp được kết nối đầy đủ hơn (với ReLU là chức năng kích hoạt của chúng) trong FC2, FC3 và FC4, dẫn đến $e_4 \in \mathbb{R}^{1\times(L_4)}$, trong đó L_4 là số đơn vị trong FC4

4.2.5. Regression Unit

Một nút duy nhất ở cuối mạng này xác định mức độ buồn ngủ bằng cách xuất ra một số thực từ 0 đến 10 tùy thuộc vào mức độ thức hoặc buồn ngủ của nháy mắt đầu vào (Phương trình 9). Thang điểm từ 0 đến 10 này giúp mạng lập mô hình chuyển đổi tự nhiên từ trạng thái tỉnh táo sang buồn ngủ không giống như các công trình trước đó [19, 20], trong đó đầu vào được phân loại trực tiếp thành các lớp khác nhau một cách riêng biệt.

$$out = 10 \times \text{Sigmoid}(\mathbf{e}_4 W_o + b_o) \qquad (9)$$

Ở đây, $w_0 \in \mathbb{R}^{L_4 \times 1}$ và $b_0 \in \mathbb{R}^{1 \times 1}$ là các tham số hồi quy, và out $\in \mathbb{R}^{1 \times 1}$ là đầu ra hồi quy cuối cùng.

4.2.6. Discretization and Voting

Khi ai đó đang buồn ngủ, điều đó không có nghĩa là tất cả những cái chớp mắt của họ nhất thiết phải đại diện cho sự buồn ngủ. Do đó, điều quan trọng là phải phân loại mức độ buồn ngủ của mỗi video là trạng thái chiếm ưu thế nhất được dự đoán từ tất cả các chuỗi nháy mắt trong video đó. Ở bước đầu tiên, chúng tôi sử dụng Eq.10 để tách biệt đầu ra hồi quy cho từng lớp được xác định trước.

$$\text{class}(out) = \begin{cases} Alert, & 0.0 \leq out < 3.3 \\ LowVigilant, & 3.3 \leq out \leq 6.6 \\ Drowsy, & 6.6 < out \leq 10 \end{cases} \qquad (10)$$

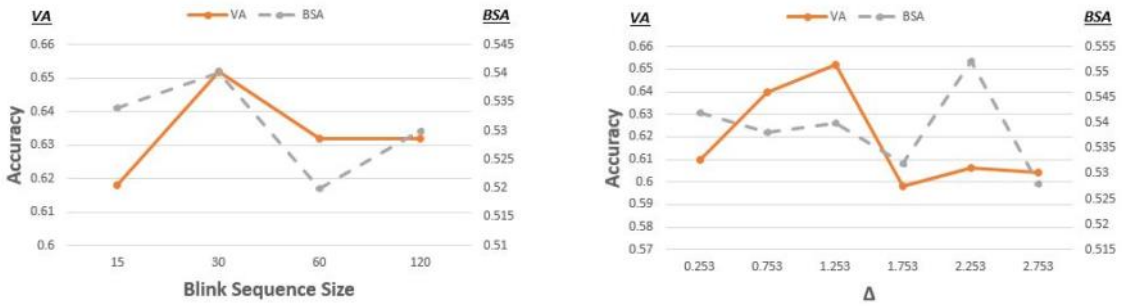
Giả sử có K nháy mắt trong video V. Sử dụng cửa sổ trượt có độ dài T, mỗi T nháy mắt liên tiếp tạo thành một chuỗi nháy mắt được đưa ra làm đầu vào cho mạng (Phương trình 7), dẫn đến có thể có nhiều chuỗi nháy mắt. Lớp được dự đoán thường xuyên nhất từ nhiều chuỗi này sẽ là kết quả phân loại cuối cùng của video V. Tác động tích cực của việc bỏ phiếu được hiển thị sau trong kết quả của chúng tôi.

4.2.7. Loss Function

Mô hình của chúng tôi học cách không phạt các dự đoán (**out**$_i$) nằm trong một khoảng cách nhất định $\sqrt{\Delta}$ của nhãn thực (t_i) cho tất cả N chuỗi huấn luyện, và thay vào đó phạt các dự đoán kém chính xác hơn bậc hai theo sai số bình phương của chúng. Kết quả là, mô hình của chúng tôi quan tâm nhiều hơn đến việc phân loại từng trình tự một cách chính xác hơn là hồi quy hoàn hảo. Thuộc tính này giúp chúng ta cùng thực hiện hồi quy và phân loại bằng cách giảm thiểu hàm mất mát sau:

$$loss = \frac{\sum_{i=1}^{N} \max(0, |out_i - t_i|^2 - \Delta)}{N} \qquad (11)$$

5. Experiments



(a) (b)

Figure 4. The effect of blink sequence size and Δ to the accuracy metrics.

Fig 4. Ảnh hưởng của kích thước chuỗi nháy mắt và Δ đến các chỉ số độ chính xác.

5.1. Evaluation Metrics

Chúng tôi đã thiết kế bốn chỉ số để đánh giá đầy đủ mô hình của chúng tôi từ các quan điểm khác nhau và ở các giai đoạn khác nhau của quá trình.

5.1.1. Blink Sequence Accuracy (BSA)

Chỉ số này đánh giá kết quả trước "giai đoạn bỏ phiếu" và sau "tùy ý" trên tất cả các chuỗi nháy mắt thử nghiệm.

5.1.2. Blink Sequence Regression Error (BSRE)

Chúng tôi định nghĩa BSRE như sau:

$$BSRE = \frac{\sum_{i=1}^{M} C_i^s |out_i - S_i|^2}{M} \qquad (12)$$

Trong phương trình trên, $\mathbf{C_i^s}$ là một giá trị nhị phân, bằng 0 nếu phân đoạn nháy mắt thứ i đã được phân loại đúng và bằng 1 nếu không. Phương thức 12 phạt mỗi trình tự nháy mắt được phân loại sai i bằng một số hạng bậc hai đối với khoảng cách của đầu ra hồi quy đến đường biên trạng thái thực gần nhất ($\mathbf{S_i}$) được xác định trong Phương trình 10. Các chuỗi nháy mắt được phân loại chính xác không gây ra lỗi BSRE.

5.1.3. Video Accuracy (VA)

“Độ chính xác của video” là chỉ số chính về độ chính xác, nó bằng phần trăm của toàn bộ video (không phải từng đoạn video riêng lẻ) đã được phân loại chính xác.

5.1.4. Video Regression Error (VRE)

VRE được định nghĩa là:

$$VRE = \frac{\sum_{j=1}^{Q} C_j^v \left| \frac{1}{K_j} \sum_{i=1}^{K_j} (out_{i,j}) - S_j \right|^2}{Q} \qquad (13)$$

Ở trên, Q là tổng số video trong tập hợp thử nghiệm và $\mathbf{C_j^v}$ là giá trị nhị phân, bằng 0 nếu video thứ j đã được phân loại đúng và bằng 1 nếu không. $\mathbf{K_j}$ là số tất cả các chuỗi nháy mắt trong video j. Các video được phân loại chính xác hoàn toàn không góp phần gây ra lỗi VRE. Đối với VA cố định, giá trị VRE cho biết biên độ sai số đối với các video được phân loại sai.

5.2. Implementation

Chúng tôi đã sử dụng một phần của tập dữ liệu RLDD làm tập thử nghiệm của mình và bốn phần còn lại để đào tạo. Sau khi lặp lại quy trình này cho mỗi lần gấp, kết quả được tính trung bình trên năm lần gấp. Đối với tham số T được xác định trong Phần 4.2, chỉ định số lần nháy mắt liên tiếp được cung cấp làm đầu vào cho mạng, chúng tôi đã sử dụng giá trị 30 (Hình 4a). Các video có ít hơn 30 lần nháy mắt không có đệm. Các chuỗi nháy mắt được tạo bằng cách áp dụng cửa sổ trượt này gồm 30 lần nháy mắt trên mỗi video, với khoảng cách là hai lần. Nếu kích thước cửa sổ quá lớn, sự phụ thuộc lâu vào các lần nháy mắt trước đó có thể trì hoãn đáng kể đầu ra chính xác trong khi chuyển đổi từ trạng thái này sang trạng thái khác.

Chúng tôi đã chú thích tất cả các chuỗi bằng nhãn của video mà chúng được lấy từ đó. Mô hình của chúng tôi đã được đào tạo trên khoảng 7000 chuỗi nháy mắt (tùy thuộc vào lần đào tạo) bằng cách sử dụng trình tối ưu hóa Adam [14] với tốc độ học tập là 0,000053, Δ là 1,253 (Hình 4b) và kích thước lô là 64 cho 80 epochs trong cả năm lần . Chúng tôi cũng sử dụng chuẩn hóa hàng loạt và chuẩn hóa L2 với hệ số (λ) là 0,1. Mô-đun HM-LSTM có bốn lớp với 32 trạng thái ẩn cho mỗi lớp. Chi tiết hơn về kiến trúc được hiển thị trong Hình 2.

5.3. Experimental Results

Model	Evaluation Metric			
	BSRE	VRE	BSA	VA
HM-LSTM network	**1.90**	**1.14**	**54%**	**65.2%**
LSTM network	3.42	2.68	52.8%	61.4%
Fully connected layers	2.85	2.17	52%	57%
Human judgment	—	2.01	—	57.8%

Table 2. This table numerically compares the performance of our model with two simplified versions of the network and human judgment using four predefined metrics. The above values are the final averaged values across all test folds.

Figure 5. Confusion matrices for: (a) our proposed model and (b) human judgment results (video accuracy).

Fig 5. Confusion matrices

Case	Metric-Fold									
	A-f1	R-f1	A-f2	R-f2	A-f3	R-f3	A-f4	R-f4	A-f5	R-f5
PM	0.64	2.42	0.61	1.04	0.70	0.58	0.64	0.85	0.67	0.81
HJ	0.62	1.37	0.59	2.3	0.60	1.96	0.53	2.32	0.55	2.07

A-f i: VA for fold i
R-f i: VRE for fold i

Table 3. Results of our Proposed Model (PM) and Human Judgment (HJ) measured by VA and VRE

Trong phần này, chúng tôi đánh giá phương pháp cơ sở của chúng tôi đối với tiêu chuẩn đánh giá của con người được giải thích trong phần 3.4. Do thiếu phương pháp hiện đại trên tập dữ liệu thực tế và công khai, chúng tôi so sánh phương pháp cơ sở của chúng tôi với hai biến thể của đường ống để cho thấy rằng toàn bộ đường ống hoạt động tốt nhất với các ô HM-LSTM. Phiên bản đầu tiên có kiến trúc giống như mạng của chúng tôi, với các ô LSTM điển hình [10] được sử dụng thay cho các ô HM-LSTM. Phiên bản thứ hai là phiên bản đơn giản hơn với cùng một kiến trúc sau khi loại bỏ mô-đun HM-LSTM, nơi chuỗi đầu vào được cấp cho mạng đa lớp được kết nối đầy đủ.

Kết quả so sánh của chúng tôi với hai biến thể này và tiêu chuẩn đánh giá của con người được liệt kê trong Bảng 2. Bảng này cho thấy kết quả xác thực chéo cuối cùng của việc phát hiện buồn ngủ theo các số liệu được xác định trước. So sánh này không chỉ làm nổi bật thông tin thời gian trong nháy mắt mà còn cho thấy độ chính xác tăng 4% mà chúng tôi đạt được sau khi chuyển sang HM-LSTM từ các ô LSTM điển hình. Như được chỉ ra bởi các chỉ số BSRE và VRE trong Bảng 2, biên độ sai số cho hồi quy cũng thấp hơn đáng kể trong mạng HM-LSTM so với hai mạng còn lại. Kết quả đối với mạng LSTM và HM-LSTM cho thấy rằng các mô hình thời gian cung cấp các giải pháp phát hiện buồn ngủ tốt hơn so với các lớp được kết nối đầy đủ đơn giản.

Như đã đề cập trước đây, tất cả các chuỗi nháy mắt trong mỗi video đều được dán nhãn giống nhau. Tuy nhiên, trên thực tế, không phải tất cả các lần chớp mắt đều thể hiện mức độ buồn ngủ như nhau. Sự khác biệt này là một lý do quan trọng khiến BSA không cao, và "biểu quyết" bù đắp điều đó dẫn đến độ chính xác cao hơn trong VA.

Hình 5a cho thấy rằng tầng lớp trung lưu (cảnh giác thấp), theo dự kiến, là tầng lớp khó phân loại nhất, nơi hầu hết bị phân loại sai vì "buồn ngủ". Mặt khác, mô hình của chúng tôi phân loại các đối tượng cảnh

giác và buồn ngủ một cách rất tự tin với độ chính xác hơn 80%, và hiếm khi phân loại sai mức độ cảnh giác cho buồn ngủ hoặc ngược lại. Điều này có nghĩa là kết quả đáng tin cậy nhất trong thực tế.

Ngoài ra, mô hình của chúng tôi phát hiện các dấu hiệu sớm và các trường hợp buồn ngủ tinh vi tốt hơn con người trong bộ dữ liệu RLDD bằng cách chỉ phân tích hành vi chớp mắt theo thời gian. Các kết quả định lượng chi tiết cho tất cả các lần gấp và giá trị trung bình cuối cùng được liệt kê trong Bảng 3 và Bảng 2 tương ứng. Mô hình phát hiện buồn ngủ của chúng tôi có khoảng 50.000 thông số có thể huấn luyện. Việc lưu trữ các thông số đó không chiếm nhiều dung lượng bộ nhớ, và do đó, mô hình có thể dễ dàng được lưu trữ trên cả điện thoại di động cấp thấp. Về thời gian chạy (ở giai đoạn đánh giá, sau khi đào tạo), hệ thống end-to-end xử lý khoảng 35-80 khung hình / giây (đối với phạm vi kích thước khung hình từ 568x320 đến 1920x1080), trên máy trạm Linux có Intel Xeon Bộ vi xử lý CPU E3-1270 V2 chạy ở tốc độ 3,5 GHz và bộ nhớ 16GB.

6. Conclusions

Trong bài báo này, chúng tôi đã trình bày một tập dữ liệu về buồn ngủ trong đời thực (RLDD) mới và công khai, theo hiểu biết của chúng tôi, lớn hơn đáng kể so với các tập dữ liệu hiện có, với gần 30 giờ video. Chúng tôi cũng đã đề xuất một phương pháp cơ bản từ đầu đến cuối sử dụng mối quan hệ thời gian giữa các lần chớp mắt để phát hiện buồn ngủ nhiều giai đoạn. Phương pháp được đề xuất có nhu cầu tính toán và lưu trữ thấp. Kết quả của chúng tôi đã chứng minh rằng phương pháp của chúng tôi vượt trội hơn khả năng phán đoán của con người trong hai thước đo được thiết kế trên tập dữ liệu RLDD.

Một chủ đề có thể thực hiện trong tương lai là thêm mạng sâu không gian để tìm hiểu các đặc điểm khác của buồn ngủ ngoài nháy mắt trong video. Nhìn chung, việc chuyển từ các tính năng thủ công sang một hệ thống học tập đầu cuối là một chủ đề thú vị, nhưng lượng dữ liệu đào tạo cần thiết vẫn chưa rõ ràng vào thời điểm này. Nhìn chung, chúng tôi hy vọng rằng bộ dữ liệu công khai được đề xuất cũng sẽ khuyến khích các nhà nghiên cứu khác làm việc để phát hiện buồn ngủ và tạo ra các kết quả bổ sung và cải thiện, có thể được trùng lặp và so sánh với nhau.

7. Supplementary Material : Blink Retrieval Algorithm

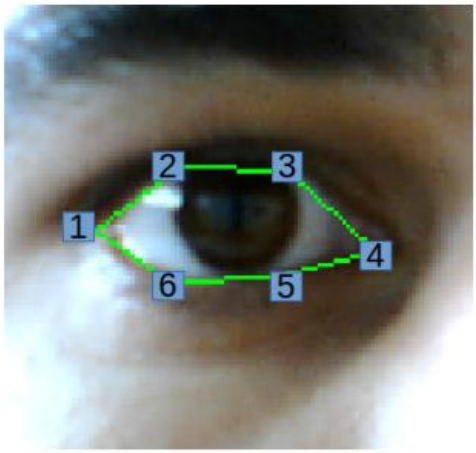


Figure 6. Six points marking each eye.

Fig 6. Sáu điểm đánh dấu mỗi mắt

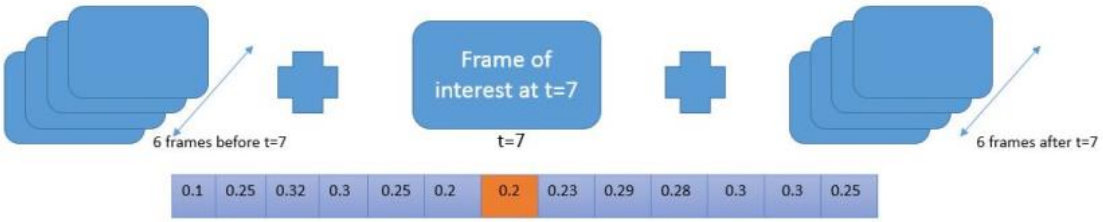


Figure 7. Presenting each frame (at t=7) by 13 numbers (EARs) concatenated from 13 frames as a feature vector.

Fig 7. Trình bày mỗi khung (tại t = 7) bằng 13 số (EAR) được ghép từ 13 khung dưới dạng một vectơ đặc trưng.

Trong các thí nghiệm của chúng tôi, chúng tôi nhận thấy rằng cách tiếp cận của Soukupova và Cech [26] thường phát hiện các chớp mắt nhanh liên tiếp chỉ trong một lần chớp mắt. Điều này gây ra vấn đề cho các bước phát hiện buồn ngủ tiếp theo vì nhiều lần chớp mắt liên tiếp thường có thể là dấu hiệu của buồn ngủ. Chúng tôi đã thêm một bước xử lý hậu kỳ ở phía trên đầu ra của [26], giúp xác định thành công nhiều lần nháy mắt có thể xuất hiện trong một lần phát hiện duy nhất do [26] tạo ra.

Theo [26], xác định EAR, cho mỗi khung, như sau:

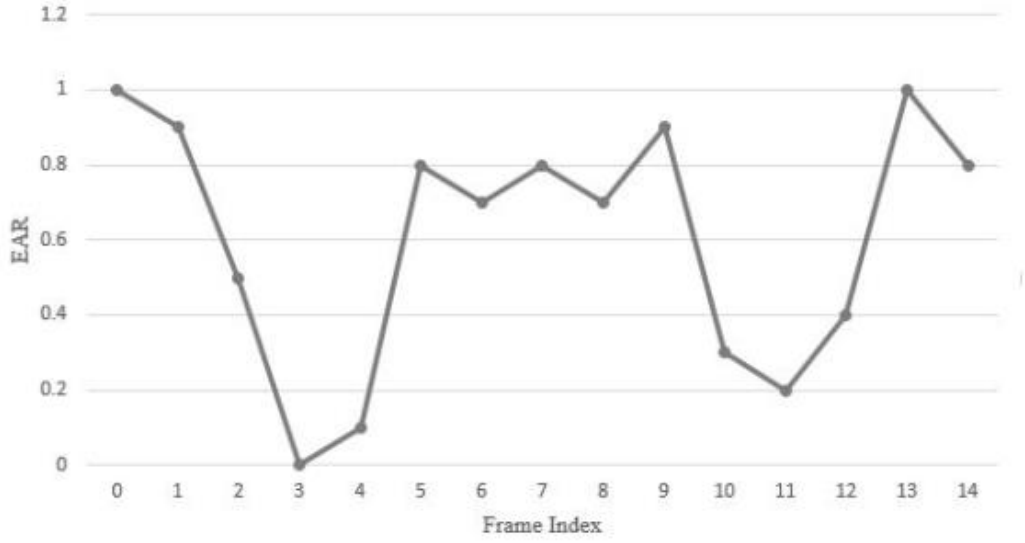
$$EAR = \frac{||\vec{p_2} - \vec{p_6}|| + ||\vec{p_3} - \vec{p_5}||}{||\vec{p_1} - \vec{p_4}||} \qquad (14)$$

Ở trên, mỗi $\overrightarrow{p_i} \in$ **{p_i , với i = 1,2,…,6}** là vị trí 2D của điểm mốc trên khuôn mặt từ vùng mắt, như được minh họa trong Hình 6. Trong [26], bộ phân loại SVM phát hiện nháy mắt mắt dưới dạng một mẫu của các giá trị EAR trong một cửa sổ thời gian ngắn có kích thước 13 được mô tả trong Hình 7. . Kích thước cửa sổ cố định này được chọn dựa trên lý do là mỗi lần nháy mắt dài khoảng 13 khung hình. Một lần nháy mắt trung bình mất khoảng 200 mili giây đến 400 mili giây [21, 24], tương đương với sáu đến mười hai khung hình cho một video được quay ở tốc độ 30 khung hình / giây. Ngay cả khi 13 khung hình là một ước tính tốt cho độ dài của một nháy mắt, thì cách tiếp cận này sẽ không xử lý các nháy mắt nhanh liên tiếp.

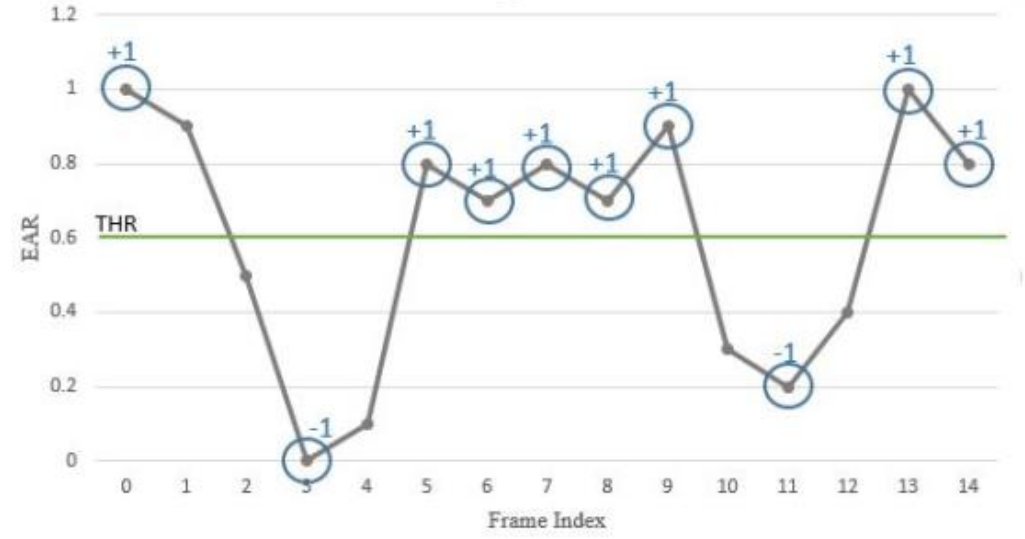
Như được mô tả trong Hình 7, mỗi giá trị trong vectơ 13 chiều này tương ứng với EAR của một khung với khung quan tâm nằm ở giữa. Bộ phân loại SVM lấy các vectơ 13D này làm đầu vào và phân loại chúng là “mở” hoặc “đóng” (được gọi cụ thể hơn là khung quan tâm trong mỗi vectơ đầu vào). Một số nhãn "đóng" liên tiếp biểu thị một nháy mắt có độ dài bằng M. Sau đó, các giá trị EAR của M khung này được lưu trữ theo thứ tự x và được đưa vào "Thuật toán truy xuất nháy mắt", như được giải thích trong Mục 1, để xử lý sau (Hình 8a). Chuỗi các giá trị EAR cho **một lần** nháy mắt [26] sẽ được coi là **ứng cử viên** cho **một hoặc nhiều lần** nháy mắt.

Thuật toán này chạy trong thời gian Θ (M), trong đó M là số khung hình trong đoạn video được sử dụng làm đầu vào cho thuật toán. Trên thực tế, thuật toán chạy trong thời gian thực. Ngoài ra, Alg.1 đặt một khung xác định về thời điểm một nháy mắt bắt đầu, kết thúc hoặc đạt đến điểm cuối của nó dựa trên điểm cực trị của tín hiệu EAR của nó. Để có kết quả tốt hơn, x được chuyển qua bộ lọc trung vị / trung bình để xóa nhiễu và sau đó được đưa vào thuật toán.

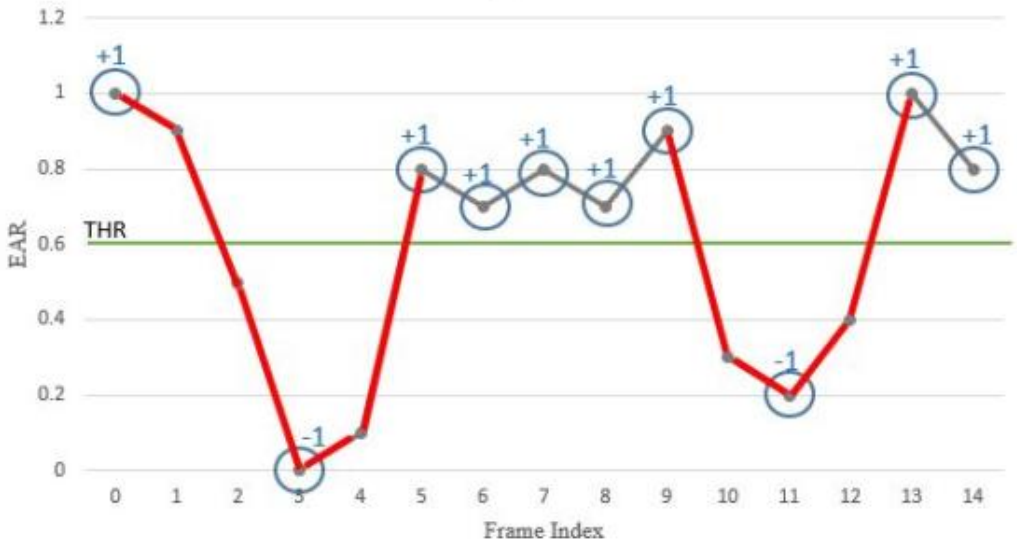
Tại bước 1, đạo hàm của x được lấy. Sau đó, các đạo hàm bằng không được sửa đổi, ở bước 2 và 3, để các đạo hàm đó có cùng dấu với đạo hàm ở bước thời gian trước đó của chúng. Việc sửa đổi này giúp tìm cực trị cục bộ, khi các điểm mà dấu đạo hàm thay đổi (bước 4 đến bước 7). Ngưỡng, được xác định ở bước 8, được sử dụng để loại bỏ các thăng trầm nhỏ trong x do nhiễu và không nháy mắt. Điểm cực trị trong x được khoanh tròn trong Hình 8b và được gắn nhãn (+1 hoặc -1) so với ngưỡng (bước 9 đến 11). Mỗi hai điểm cực viễn liên tiếp là dấu hiệu của chuyển động đi xuống hoặc hướng lên của mắt trong nháy mắt nếu hai điểm đó được nối với nhau để liên kết hoặc các liên kết giữa chúng vượt qua đường ngưỡng (bước 12 và 13). Hình 8c làm nổi bật các liên kết này bằng màu đỏ. Cuối cùng, mỗi lần ghép nối các liên kết màu đỏ này tương ứng với một lần nháy mắt với các điểm bắt đầu, điểm kết thúc và điểm cuối như được mô tả trong Hình 8d (bước 14 đến cuối).


1.2
1
0.8
0.6
0.4
0.2
0
EAR
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Frame Index

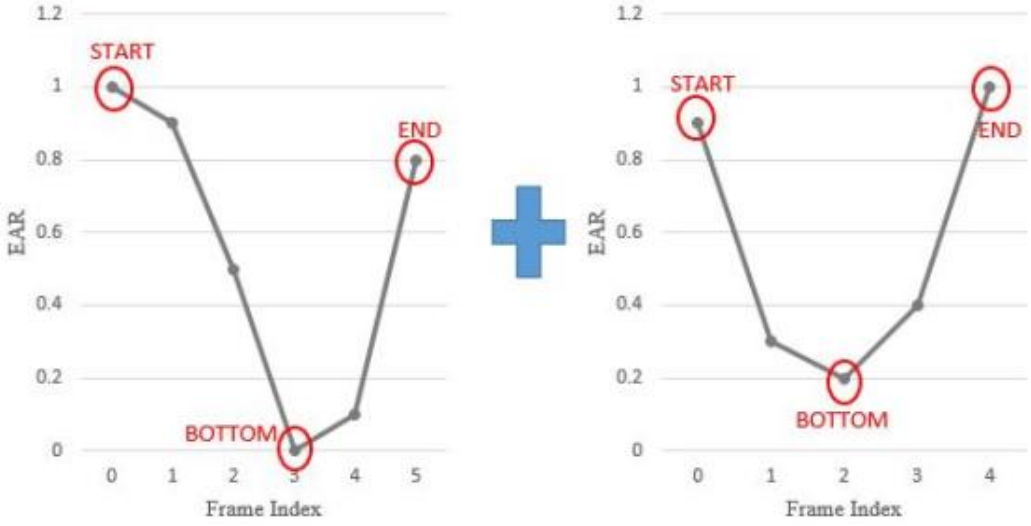
(a)


1.2
1
0.8
0.6
0.4
0.2
0
EAR
THR
+1
-1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Frame Index

(b)


1.2
1
0.8
0.6
0.4
0.2
0
EAR
THR
+1
+1
+1
+1
+1
+1
+1
+1
-1
-1
0 1 2 3 4 5 6 7 8 9 10 11 12 13 14
Frame Index

(c)


1.2
1
0.8
0.6
0.4
0.2
0
EAR
START
END
BOTTOM
0 1 2 3 4 5
Frame Index
1.2
1
0.8
0.6
0.4
0.2
0
EAR
START
END
BOTTOM
0 1 2 3 4
Frame Index

(d)

(d)

Figure 8. The Blink Retrieval Algorithm steps: (a) $\mathbf{x}$ with size $M = 15$ as the input for Alg. 1. (b) The indices of circled points form $\mathbf{e}$, and the set of +1 and -1 labels forms $\mathbf{t}$ with $P = 8$. (c) The red lines indicate where $\mathbf{z}$ values are negative. (d) Two ($N = 2$) blinks are retrieved with definite start, end and bottom points.

Fig 8. Các bước của Thuật toán Truy xuất Nháy mắt: (a) x với kích thước M = 15 làm đầu vào cho Alg. 1. (b) Chỉ số của các điểm được khoanh tròn tạo thành e, và tập hợp các nhãn +1 và -1 tạo thành t với P = 8. (c) Các đường màu đỏ cho biết các giá trị z là âm. (d) Hai (N = 2) lần nháy mắt được truy xuất với các điểm bắt đầu, điểm kết thúc và điểm cuối cùng xác định.

Algorithm 1 Blink Retrieval Algorithm

Input The initial detected EAR signal $\mathbf{x} \in \mathbb{R}^M$, where M is the size of the $\mathbf{x}$ time series, as a candidate for one or more blinks and epsilon=0.01

Output N retrieved blinks, $N \ll M$

1: $\dot{\mathbf{x}}[n] \leftarrow \mathbf{x}[n+1] - \mathbf{x}[n], \forall\, n \in \{i | i = 0, 1, ..., M-2\}$
2: **if** $\dot{\mathbf{x}}[0] = 0$ **then** $\dot{\mathbf{x}}[0] \leftarrow -1 \times$ epsilon
3: $\dot{\mathbf{x}}[n] \leftarrow \dot{\mathbf{x}}[n-1] \times$ epsilon, $\forall\, n \in \{i | \dot{\mathbf{x}}[i] = 0 \wedge i \neq 0\}$ to avoid zero derivatives for steps 4 and 6
4: $\mathbf{c}[n] \leftarrow \dot{\mathbf{x}}[n+1] \times \dot{\mathbf{x}}[n], \forall\, n \in \{i | i = 0, 1, ..., M-3\}$
5: Define $\mathbf{e} \in \mathbb{R}^{P+2}, P \leq M-2$ to store the indices for the P extrema, the first and the last points in $\mathbf{x}$
6: $\mathbf{e}[0] \leftarrow 0$, $\mathbf{e}[P+1] \leftarrow M-1$, supposing the first and last points in $\mathbf{x}$ are maxima
7: $\mathbf{e}[k] \leftarrow n+1, \forall\, (n \in \{i | \mathbf{c}[i] < 0\} \wedge k \in \{i | i = 1, 2, ..., P\})$ ▷ Indices of P+2 extrema, including the first and last points in $\mathbf{x}$ are stored in order
8: Define $THR \leftarrow 0.6 \times \max(\mathbf{x}) + 0.4 \times \min(\mathbf{x})$, as a threshold

Define $\mathbf{t} \in \mathbb{R}^{P+2}$, to store +1 or -1 for extrema above and below threshold respectively

$\mathbf{t}[0] \leftarrow +1, \mathbf{t}[P+1] \leftarrow +1$, supposing the first and last points in $\mathbf{x}$ are maxima

Append +1 in $\mathbf{t}$ for each $n \in \{i|\mathbf{x}[\mathbf{e}[i]] > THR\}$, and append -1 in $\mathbf{t}$ for each $n \in \{i|\mathbf{x}[\mathbf{e}[i]] \leq THR\}$, all in the order of the indices in $\mathbf{e}$

Define $\mathbf{z} \in \mathbb{R}^{P+1}$, $\mathbf{z}[n] \leftarrow \mathbf{t}[n+1] \times \mathbf{t}[n]$

Define $\mathbf{s}$, to store the indices of all negative values in $\mathbf{z}$, representing the downward and upward movements of eyes in a blink

$N \leftarrow \frac{length(\mathbf{s})}{2}$ ▷ N is the number of sub blinks, and $length(\mathbf{s})$ is always an even number

for $i \leftarrow 0$ to $N-1$ **do** ▷ Define for $blink_i$:

    StartIndex $\leftarrow \mathbf{e}[\mathbf{s}[2 \times i]]$,

    EndIndex$\leftarrow \mathbf{e}[\mathbf{s}[2 \times i + 1] + 1]$,

    BottomIndex$\leftarrow \mathbf{e}[\mathbf{s}[2 \times i + 1]]$

return start, end and bottom points of the N retrieved blinks in $\mathbf{x}$