Tóm Tắt Video Bằng Mạng Học Sâu Nơ-ron

1st Chung Kim Khánh  
Đại học Khoa học tự nhiên (ĐHQG)Hồ Chí Minh, Việt Nam  
ckkhanh19@clc.fitus.edu.vn

2nd Phạm Nguyễn Anh Quốc  
Đại học Khoa học tự nhiên (ĐHQG)Hồ Chí Minh, Việt Nam  
pnaquoc19@clc.fitus.edu.vn

*Tóm lược*—Công nghệ tóm tắt video dùng để tạo ra bản video mới tóm tắt một cách ngắn gọn nhưng diễn đạt đầy đủ thông tin của video ban đầu. Vì thế, ta thực hiện chọn lọc những đoạn video có thông tin quan trọng để sử dụng. Một số phương pháp đã được đưa ra và cải tiến qua nhiều thập kỷ. Hiện tại, các phương pháp đa phần đều dựa trên cơ sở mô hình kiến trúc mạng nơ-ron. Trong tài liệu này, chúng tôi tập trung vào những cải tiến đã có ở hiện tại và khảo sát toàn diện về các phương pháp tóm tắt video dựa trên mạng học sâu. Sau khi trình bày động lực phát triển của công nghệ tóm tắt video, chúng tôi sẽ đưa ra giải thuật và thảo luận về các đặc trưng của giải thuật dựa trên mạng học sâu điển hình. Tiếp theo, chúng tôi sẽ đề xuất phân loại các thuật toán hiện có và phân tích có cơ sở dựa trên những tài liệu liên quan để đưa ra đề xuất cho các phát triển ở trong tương lai. Cuối cùng, chúng tôi đánh giá khách quan các thuật toán tóm tắt video và so sánh hiệu suất của các phương pháp dựa trên mạng học sâu. Từ kết quả so sánh này, cũng như những ý kiến đã được đưa ra trước đó về dữ liệu và tính phù hợp của giao thức đánh giá, chúmg tôi sẽ đưa ra các hướng nghiên cứu tiềm năng của các thuật toán trong tương lai.

Từ khoá—Tóm tắt video, Mạng học sâu nơ-ron, Học có giám sát, Học không có giám sát, Tóm tắt dữ liệu, Giao thức đánh giá

# Giới Thiệu

Youtube được biết đến là một nền tảng chia sẻ video trực tuyến của Mỹ được sử dụng phổ biến rộng rãi trên khắp thế giới. Tháng 7 năm 2015, Youtube công bố rằng họ nhận được hơn 400 giờ nội dung video trên mỗi phút1. Điều này tương đương với 65.7 năm nội dung được tải lên mỗi ngày. Tuy nhiên, Youtube chỉ là một trong các nền tảng chia sẻ video trên thế giới (Tiktok, DailyMotion,…). Ngoài ra còn có các mạng xã hội (Facebook, Instagram, Weibo,…), kho lưu trữ trực tuyến của các tổ chức truyền thông và tin tức đang lưu trữ một khối lượng lớn các nội dung video. Không dừng lại ở đó, các nhà phát triển không ngừng hướng đến nhu cầu của người tiêu dùng hiện tại trên các nền tảng và các thiết bị quay phim trực tuyến (điện thoại thông minh, đồng hồ thông minh,…) bằng cách trang bị cảm biến mạnh mẽ tăng chất lượng video và cho phép tải video đã quay lên Web ngay lập tức. Vì vậy, làm cách nào để người dùng có thể tìm kiếm thông tin một cách nhanh chóng, hiệu quả cao và chính xác giữa vô vàn thông tin như thế này? Câu trả lời cho câu hỏi này không chỉ liên quan đến việc truy xuất thông tin từ video mà còn liên quan đến công nghệ tóm tắt video. Việc tóm tắt giúp cho thông tin được ngắn gọn hơn nhưng mang nội dung trọng tâm của toàn bộ video. Giữa những nội dung video trên Web, việc tóm tắt video hiệu quả sẽ tạo điều kiện thuận lợi cho người phê duyệt và đưa video vào đúng nhóm nội dung của nó. Qua đó tăng lượt tương tác của người xem và mức độ tiêu thị nội dung bằng cách hướng đến đúng đối tượng tiềm năng.

Ứng dụng tóm tắt video được sử dụng trong rất nhiều lĩnh vực. Cụ thể ở các tổ chức truyền thông đa phương tiện, họ sử dụng công nghệ này để chỉ mục, phê duyệt, truy xuất và quảng bá một cách hiệu quả các nội dung video. Từ đó, phân loại video phù hợp với các nhóm đối tượng có nhu cầu. Việc làm này giúp tăng lượt xem cũng như lượng khách hàng tiềm năng tiếp cận sản phẩm để phục vụ cho quảng cáo. Đồng nghĩa với việc tăng lợi nhuận, lượng tiêu thụ và doanh thu cho sản phẩm quảng cáo. Không dừng lại ở lĩnh vực truyền thông, tóm tắt video còn hỗ trợ trong việc tạo ra những bộ phim giới thiệu (trailers) và các tập ngắn của phim truyền hình dài tập. Đặc biệt hơn, tóm tắt video giúp tóm tắt lại nội dung chính đã diễn ra trong ngày hoặc trong tuần của camera giám sát, giúp cho việc giám sát tiến độ và bảo mật hiệu quả hơn về mặt thời gian.

Một số khảo sát về tóm tắt đã xuất hiện trong tài liệu. Một trong những tác phẩm đầu tiên phải kể đến Barbieri và cộng sự (2003) [1] – phân loại các kịch bản của video, đặc điểm của cách tiếp cận với việc tóm tắt video. Trong một nghiên cứu ban đầu khác, Li và cộng sự (2006) [2] đã chia các phương pháp tiếp cận tóm tắt hiện có thành các phương pháp dựa trên tiện ích sử dụng các mô hình chú ý để xác định các đối tượng và cảnh nổi bật, và các phương pháp dựa trên cấu trúc xây dựng dựa trên các cảnh quay và video. Truong và cộng sự (2007) [3] thì lại thảo luận về các thuộc tính ảnh hưởng đến kết quả trong quá trình tóm tắt (miền video, mức độ chi tiết của các đoạn video được sử dụng, phương pháp được sử dụng và loại tóm tắt). Money và cộng sự (2008) [4] phân loại phương pháp thành phương pháp dựa trên phân tích luồng video, phương pháp xử lý siêu dữ liệu video theo ngữ cảnh và cuối cùng là phương pháp kết hợp cả hai. Jiang và cộng sự (2009) [5] thảo luận về một số cách tiếp cận tóm tắt video đặc trưng, bao gồm việc trích xuất các đặc điểm hình ảnh cấp thấp để đánh giá độ tương đồng của khung hình hoặc thực hiện lựa chọn khung hình chính dựa trên phân cụm; phát hiện các sự kiện chính của video bằng cách sử dụng bộ mô tả chuyển động; và xác định cấu trúc video bằng cách sử dụng Eigen-features. Hu và cộng sự (2011) [6] phân loại các phương pháp tóm tắt thành phương pháp nhắm mục tiêu dư thừa trực quan tối thiểu, phương pháp dựa vào phát hiện đối tượng hoặc sự kiện và phương pháp khác dựa trên tích hợp đa phương thức. Ajmal và cộng sự (2012) [7] đã dựa trên các tài liệu liên quan phân loại thành phương pháp dựa trên phân nhóm, phương pháp dựa trên việc phát hiện các sự kiện chính của câu chuyện,… Tuy nhiên, tất cả các tác phẩm nói trên (xuất bản từ năm 2003 đến 2012) đều báo cáo về các cách tiếp cận ban đầu đối với việc tóm tắt video; họ không trình bày bối cảnh tóm tắt đã phát triển như thế nào trong những năm qua và đặc biệt là sau khi giới thiệu các thuật toán học sâu.

Trong các nghiên cứu gần đây hơn của Molino và cộng sự (2017) [8] tập trung chủ yếu vào việc tóm tắt trọng tâm của video và thảo luận về các thông số kỹ thuật cũng như các thách thức. Trong một số công trình nghiên cứu gần đây, Basavarajaiah và cộng sự (2019) [9] đưa ra phân loại các phương pháp tiếp cận tóm tắt khác nhau, bao gồm một số phương pháp dựa trên học sâu được đề cập gần đây. Tuy nhiên, họ chủ yếu tập trung vào các thuật toán tóm tắt có thể áp dụng được trên miền nén. Cuối cùng là cuộc khảo sát của Vivekraj và cộng sự (2019) [10], họ trình bày dựa trên phân loại hai chiều, liên quan đến các phương thức dữ liệu được sử dụng trong quá trình phân tích và kết hợp các khía cạnh của con người. Đối với phương pháp thứ hai, nó tiếp tục chia ra thành phương pháp tạo tóm tắt bằng cách mô hình hóa sự hiểu biết và sở thích của con người (ví dụ là sử dụng các mô hình chú ý, ngữ nghĩa của nội dung trực quan hoặc các chú thích về sự thật cơ bản và các thuật toán học máy), và phương pháp tiếp cận thông thường dựa vào việc xử lý thống kê các tính năng cấp thấp của video. Tuy nhiên, không có cuộc khảo sát nào ở trên trình bày một cách toàn diện những phát triển hiện tại của việc tóm tắt video dựa trên mạng học sâu nơ-ron nâng cao. Trên thực tế, lĩnh vực nghiên cứu này rất hay vì một số phương pháp tiếp cận mới đang được trình bày hằng năm trên các tạp chí và hội nghĩ quốc tế đều được đánh giá cao. Trong cuộc khảo sát này, chúng tôi nghiên cứu chi tiết hơn 40 thuật toán tóm tắt video dựa trên phương pháp học sâu khác nhau được đề xuất trong 5 năm qua. Ngoài ra, so sánh hiệu suất tóm tắt video đã được báo cáo trong các phương pháp tiếp cận dựa trên học sâu gần đây với hiệu suất trong các phương pháp tiếp cận khác [10]-[13] cho thấy rằng trong hầu hết các trường hợp, các phương pháp dựa trên học sâu vượt trội hơn đáng kể so với các phương pháp truyền thống. Qua đó, chúng tôi muốn tiếp tục đóng góp vào công trình phát triển tóm tắt video dựa trên học sâu bằng cách trình bày và thảo kuaajn về các khía cạnh liên quan chẳng hạn như các giao thức được sử dụng để đánh giá.

Bố cục của tài liệu này như sau: Phần II mở đầu bài viết này bằng cách xác định các vấn đề trong tóm tắt video tự động và trình bày các kiểu tóm tắt video nổi bật. Sau đó, chúng tôi sẽ chung cấp mô tả chi tiết hơn về quy trình phân tích của các thuật toán tóm tắt dựa trên học sâu và giới thiệu phân loại theo phương thức dữ liệu được sử dụng, chiến lược đào tạo được chấp nhận và phương pháp tiếp cận học sâu đã được triển khai. Cuối cùng là thảo luận về các khía cạnh liên quan đến phương pháp tóm tắt đã đề cập, chẳng hạn như các thuộc tính mông muốn của bản tóm tắt video tĩnh (dựa trên khung hình) và thời lượng của bản động (dựa trên phân đoạn).

# Các Vấn Đề Liên Quan

**Video summarization** nhằm mục đích tạo bản tóm tắt ngắn nội dung video bằng cách chọn các phần quan trọng và nhiều thông tin nhất của nó. Bản tóm tắt được tạo thường bao gồm một tập hợp các video **key-frames** (khung hình chính) hoặc các **key-fragments** (đoạn video chính) đã được ghép theo thứ tự thời gian để tạo thành một video ngắn hơn.

Loại trước đây của tóm tắt video được gọi là **video storyboard** (bảng phân cảnh video) và loại sau đó được gọi là **video skim** (video lướt). Một lợi thế của video skim so với các bộ frame tĩnh là khả năng bao gồm các yếu tố âm thanh và chuyển động mang lại khả năng tường thuật và biểu cảm, cũng như lượng thông tin được truyền tải cao hơn. Hơn nữa, người xem thường cảm thấy thú vị khi xem một video skim hơn là một slide show gồm các khung hình. Mặt khác, **storyboard** không bị giới hạn về thời gian hoặc đồng bộ hóa, và do đó, chúng mang lại sự linh hoạt hơn về mặt tổ chức dữ liệu cho các mục đích duyệt và điều hướng.

Sự biểu diễn ở level cao của một quy trình tóm tắt video dựa trên nền tảng deep-learning điển hình được mô tả trong Hình 1.

Diagram

Description automatically generated

*Hình 1. Trình bày cấp cao về quy trình phân tích của các phương pháp tóm tắt video dựa trên deep-learning để tạo video storyboard và video skim.*

Bước đầu tiên của phân tích liên quan đến việc trình bày nội dung trực quan của video với sự trợ giúp của **vectơ đặc trưng**. Thông thường nhất, các vectơ như vậy được trích xuất ở frame-level, cho tất cả các frame hoặc cho một tập hợp con của chúng được chọn thông qua chiến lược lấy mẫu frame (**frame-sampling strategy**) (vd: xử lý 2 khung hình mỗi giây). Bằng cách này, các vectơ đặc trưng được trích xuất lưu trữ thông tin ở mức rất chi tiết và nắm bắt động lực của nội dung trực quan có ý nghĩa cao khi chọn các phần video tạo thành bản tóm tắt.

Thông thường, trong hầu hết các kỹ thuật tóm tắt video dựa trên deep-learning, nội dung trực quan của các khung video được thể hiện bằng các vectơ đặc trưng sâu được trích xuất với sự trợ giúp của mạng nơ-ron được training trước (**pre-trained neural networks**). Vì lý do này, nhiều loại **Convolutional Neural Network** (**CNN**) và **Deep Convolutional Neural Network** (**DCNN**) đã được sử dụng trong danh mục, bao gồm GoogleNet (Inception V1), Inception V3, AlexNet, các biến thể của ResNet và các biến thể của VGGnet. Tuy nhiên, **GoogleNet** dường như được sử dụng phổ biến nhất cho đến nay.

Sau đó, các đặc trưng được trích xuất sẽ được sử dụng bởi một mạng tóm tắt sâu (**Deep Summarizer Network**), mạng này được train bằng cách cố gắng giảm thiểu một **objective function** hoặc một tập hợp các objective function.

Đầu ra của **Deep Summarizer Network** sau khi được train có thể là một tập hợp các **key-frame** tạo thành bảng phân cảnh video tĩnh **(static video storyboard**) hoặc một tập hợp các video **key-fragments** được nối với nhau theo thứ tự thời gian và tạo thành một **short video skim**.

Đối với video storyboard, phải tương tự với bộ key-frames do con người lựa chọn và phải thể hiện sự dư thừa hình ảnh tối thiểu. Đối với phần video skim, thường phải bằng hoặc nhỏ hơn độ dài *L* được xác định trước.

Với mục đích thử nghiệm và so sánh, điều này thường được đặt

L = p · T

*trong đó, T là thời lượng video gốc, p là tỷ lệ "phần tóm tắt"/"thời lượng video gốc" (p = 0.15 là giá trị điển hình, nghĩa là phần tóm tắt không quá 15% video gốc)*

Lưu ý, một video skim yêu cầu các segmentation phải liên tục và không chồng chéo, thống nhất về mặt hình ảnh và thời gian ⭢ cung cấp một bản trình bày liền mạch của một phần câu chuyện. Dựa trên segmentation đó và ước tính **importance scores** (điểm số tầm quan trọng) của các frame bởi **Trained Deep Summarizer Network**, điểm số ở video-segment-level được tính bằng cách tính trung bình điểm số của các frame nằm trong mỗi video segment.

Sau đó, các điểm số ở segment-level này được sử dụng để chọn các key-fragments có độ dài tóm tắt *L* và hầu hết các phương pháp giải quyết bước này bằng cách giải quyết vấn đề Knapsack.

Liên quan đến loại dữ liệu được sử dụng, danh sách hiện tại về **video summarization** dựa trên deep-learning có thể được chia thành:

* Các phương pháp tiếp cận đơn phương (**unimodal approache**) chỉ sử dụng phương thức trực quan của video để trích xuất đặc trưng và tìm hiểu tóm tắt theo cách có giám sát (nhưng yếu) hoặc không giám sát.
* Các phương pháp đa phương (**Multimodal method**) khai thác siêu dữ liệu văn bản có sẵn và tìm hiểu tóm tắt theo ngữ nghĩa/danh mục theo cách được giám sát bằng cách tăng mức độ liên quan giữa ngữ nghĩa của bản tóm tắt và ngữ nghĩa của siêu dữ liệu hoặc danh mục video được liên kết

Liên quan đến chiến lược training đã được thông qua, các thuật toán video summarization dựa trên deep-learning có thể được phân loại thô trong các loại sau:

* Các phương pháp tiếp cận được giám sát (**Supervised approache**) dựa trên tập dữ liệu có chú thích sự thật cơ bản do con người gắn nhãn (dưới dạng video tóm tắt, như trong trường hợp của tập dữ liệu SumMe hoặc ở dạng importance scores ở frame-level, như trong trường hợp của tập dữ liệu TVSum [56]), dựa vào đó họ cố gắng khám phá tiêu chí cơ bản cho việc lựa chọn frame/fragment và tóm tắt video.
* Các phương pháp tiếp cận không giám sát (**Unsupervised approache**) khắc phục được nhu cầu về dữ liệu chân thực (mà quá trình sản xuất đòi hỏi thời gian và quy trình chú thích thủ công tốn nhiều công sức), dựa trên cơ chế học hỏi chỉ yêu cầu một bộ sưu tập đủ lớn các video gốc cho quá trình training của họ.
* Các phương pháp tiếp cận có giám sát yếu (**Weakly-supervised approache**), tương tự như các phương pháp tiếp cận không được giám sát, nhằm mục đích giảm bớt nhu cầu về các bộ dữ liệu được dán nhãn thủ công. Các nhãn yếu ít tốn kém hơn được sử dụng với sự hiểu biết rằng chúng không hoàn hảo so với toàn bộ các chú thích của con người, nhưng vẫn có thể được sử dụng để tạo ra các mô hình dự đoán mạnh mẽ.

Dựa trên các phân loại được mô tả ở trên, phân loại chi tiết hơn của các thư mục tham khảo liên quan được mô tả trong Hình 2. Layer áp chót của minh họa cây thực vật này cho thấy các phương pháp leaning khác nhau đã được áp dụng. Các lá của mỗi node của layer này hiển thị các kỹ thuật được sử dụng để thực hiện từng cách tiếp cận learning và chứa các tham chiếu đến các công trình có liên quan nhất trong thư mục thư khảo. Bảng phân loại này sẽ là cơ sở để trình bày các thư mục tham khảo liên quan trong phần sau.

Diagram

Description automatically generated

# Cách Tiếp Cận Học Sâu

Before you begin to format your paper, first write and save the content as a separate text file. Complete all content and organizational editing before formatting. Please note sections A-D below for more information on proofreading, spelling and grammar.

Keep your text and graphic files separate until after the text has been formatted and styled. Do not use hard tabs, and limit use of hard returns to only one return at the end of a paragraph. Do not add any kind of pagination anywhere in the paper. Do not number text heads-the template will do that for you.

## Kiến thức cơ bản về học sâu

Define abbreviations and acronyms the first time they are used in the text, even after they have been defined in the abstract. Abbreviations such as IEEE, SI, MKS, CGS, sc, dc, and rms do not have to be defined. Do not use abbreviations in the title or heads unless they are unavoidable.

## Tóm tắt video có giám sát

* Use either SI (MKS) or CGS as primary units. (SI units are encouraged.) English units may be used as secondary units (in parentheses). An exception would be the use of English units as identifiers in trade, such as “3.5-inch disk drive”.
* Avoid combining SI and CGS units, such as current in amperes and magnetic field in oersteds. This often leads to confusion because equations do not balance dimensionally. If you must use mixed units, clearly state the units for each quantity that you use in an equation.
* Do not mix complete spellings and abbreviations of units: “Wb/m2” or “webers per square meter”, not “webers/m2”. Spell out units when they appear in text: “. . . a few henries”, not “. . . a few H”.

Identify applicable funding agency here. If none, delete this text box.

* Use a zero before decimal points: “0.25”, not “.25”. Use “cm3”, not “cc”. (*bullet list*)

## Equations

The equations are an exception to the prescribed specifications of this template. You will need to determine whether or not your equation should be typed using either the Times New Roman or the Symbol font (please no other font). To create multileveled equations, it may be necessary to treat the equation as a graphic and insert it into the text after your paper is styled.

Number equations consecutively. Equation numbers, within parentheses, are to position flush right, as in (1), using a right tab stop. To make your equations more compact, you may use the solidus ( / ), the exp function, or appropriate exponents. Italicize Roman symbols for quantities and variables, but not Greek symbols. Use a long dash rather than a hyphen for a minus sign. Punctuate equations with commas or periods when they are part of a sentence, as in:

*a**b* 

Note that the equation is centered using a center tab stop. Be sure that the symbols in your equation have been defined before or immediately following the equation. Use “(1)”, not “Eq. (1)” or “equation (1)”, except at the beginning of a sentence: “Equation (1) is . . .”

## Some Common Mistakes

* The word “data” is plural, not singular.
* The subscript for the permeability of vacuum **0, and other common scientific constants, is zero with subscript formatting, not a lowercase letter “o”.
* In American English, commas, semicolons, periods, question and exclamation marks are located within quotation marks only when a complete thought or name is cited, such as a title or full quotation. When quotation marks are used, instead of a bold or italic typeface, to highlight a word or phrase, punctuation should appear outside of the quotation marks. A parenthetical phrase or statement at the end of a sentence is punctuated outside of the closing parenthesis (like this). (A parenthetical sentence is punctuated within the parentheses.)
* A graph within a graph is an “inset”, not an “insert”. The word alternatively is preferred to the word “alternately” (unless you really mean something that alternates).
* Do not use the word “essentially” to mean “approximately” or “effectively”.
* In your paper title, if the words “that uses” can accurately replace the word “using”, capitalize the “u”; if not, keep using lower-cased.
* Be aware of the different meanings of the homophones “affect” and “effect”, “complement” and “compliment”, “discreet” and “discrete”, “principal” and “principle”.
* Do not confuse “imply” and “infer”.
* The prefix “non” is not a word; it should be joined to the word it modifies, usually without a hyphen.
* There is no period after the “et” in the Latin abbreviation “et al.”.
* The abbreviation “i.e.” means “that is”, and the abbreviation “e.g.” means “for example”.

An excellent style manual for science writers is [7].

# Đánh Giá Tóm Tắt Video

After the text edit has been completed, the paper is ready for the template. Duplicate the template file by using the Save As command, and use the naming convention prescribed by your conference for the name of your paper. In this newly created file, highlight all of the contents and import your prepared text file. You are now ready to style your paper; use the scroll down window on the left of the MS Word Formatting toolbar.

## Authors and Affiliations

**The template is designed for, but not limited to, six authors.** A minimum of one author is required for all conference articles. Author names should be listed starting from left to right and then moving down to the next line. This is the author sequence that will be used in future citations and by indexing services. Names should not be listed in columns nor group by affiliation. Please keep your affiliations as succinct as possible (for example, do not differentiate among departments of the same organization).

### For papers with more than six authors: Add author names horizontally, moving to a third row if needed for more than 8 authors.

### For papers with less than six authors: To change the default, adjust the template as follows.

#### Selection: Highlight all author and affiliation lines.

#### Change number of columns: Select the Columns icon from the MS Word Standard toolbar and then select the correct number of columns from the selection palette.

#### Deletion: Delete the author and affiliation lines for the extra authors.

## Identify the Headings

Headings, or heads, are organizational devices that guide the reader through your paper. There are two types: component heads and text heads.

Component heads identify the different components of your paper and are not topically subordinate to each other. Examples include Acknowledgments and References and, for these, the correct style to use is “Heading 5”. Use “figure caption” for your Figure captions, and “table head” for your table title. Run-in heads, such as “Abstract”, will require you to apply a style (in this case, italic) in addition to the style provided by the drop down menu to differentiate the head from the text.

Text heads organize the topics on a relational, hierarchical basis. For example, the paper title is the primary text head because all subsequent material relates and elaborates on this one topic. If there are two or more sub-topics, the next level head (uppercase Roman numerals) should be used and, conversely, if there are not at least two sub-topics, then no subheads should be introduced. Styles named “Heading 1”, “Heading 2”, “Heading 3”, and “Heading 4” are prescribed.

## Figures and Tables

#### Positioning Figures and Tables: Place figures and tables at the top and bottom of columns. Avoid placing them in the middle of columns. Large figures and tables may span across both columns. Figure captions should be below the figures; table heads should appear above the tables. Insert figures and tables after they are cited in the text. Use the abbreviation “Fig. 1”, even at the beginning of a sentence.

1. Table Type Styles

| Table Head | Table Column Head | | |
| --- | --- | --- | --- |
| Table column subhead | Subhead | Subhead |
| copy | More table copya |  |  |

1. Sample of a Table footnote. (*Table footnote*)
2. Example of a figure caption. (*figure caption*)

Figure Labels: Use 8 point Times New Roman for Figure labels. Use words rather than symbols or abbreviations when writing Figure axis labels to avoid confusing the reader. As an example, write the quantity “Magnetization”, or “Magnetization, M”, not just “M”. If including units in the label, present them within parentheses. Do not label axes only with units. In the example, write “Magnetization (A/m)” or “Magnetization {A[m(1)]}”, not just “A/m”. Do not label axes with a ratio of quantities and units. For example, write “Temperature (K)”, not “Temperature/K”.

##### Acknowledgment *(Heading 5)*

The preferred spelling of the word “acknowledgment” in America is without an “e” after the “g”. Avoid the stilted expression “one of us (R. B. G.) thanks ...”. Instead, try “R. B. G. thanks...”. Put sponsor acknowledgments in the unnumbered footnote on the first page.

##### References

1. M. Barbieri, L. Agnihotri, and N. Dimitrova, “Video summarization: methods and landscape,” in Internet Multimedia Management Systems IV, J. R. Smith, S. Panchanathan, and T. Zhang, Eds., vol. 5242, International Society for Optics and Photonics. SPIE, 2003, pp. 1 – 13.
2. Ying Li, Shih-Hung Lee, Chia-Hung Yeh, and C.-J. Kuo, “Techniques for movie content analysis and skimming: tutorial and overview on video abstraction techniques,” IEEE Signal Processing Magazine, vol. 23, no. 2, pp. 79–89, 2006.
3. B. T. Truong and S. Venkatesh, “Video Abstraction: A Systematic Review and Classification,” ACM Trans. Multimedia Comput. Commun. Appl., vol. 3, no. 1, p. 3–es, Feb. 2007.
4. A. G. Money and H. Agius, “Video summarisation: A conceptual framework and survey of the state of the art,” Journal of Visual Communication and Image Representation, vol. 19, no. 2, pp. 121 – 143, 2008.
5. R. M. Jiang, A. H. Sadka, and D. Crookes, Advances in Video Summarization and Skimming. Berlin, Heidelberg: Springer Berlin Heidelberg, 2009, pp. 27–50.
6. W. Hu, N. Xie, L. Li, X. Zeng, and S. Maybank, “A Survey on Visual Content-Based Video Indexing and Retrieval,” IEEE Trans. on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 41, no. 6, pp. 797–819, 2011.
7. M. Ajmal, M. H. Ashraf, M. Shakir, Y. Abbas, and F. A. Shah, “Video Summarization: Techniques and Classification,” in Computer Vision and Graphics, L. Bolc, R. Tadeusiewicz, L. J. Chmielewski, and K. Wojciechowski, Eds. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012, pp. 1–13.
8. A. G. del Molino, C. Tan, J. Lim, and A. Tan, “Summarization of Egocentric Videos: A Comprehensive Survey,” IEEE Trans. on HumanMachine Systems, vol. 47, no. 1, pp. 65–76, Feb 2017.
9. M. Basavarajaiah and P. Sharma, “Survey of Compressed Domain Video Summarization Techniques,” ACM Computing Surveys, vol. 52, no. 6, Oct. 2019.
10. V. V. K., D. Sen, and B. Raman, “Video Skimming: Taxonomy and Comprehensive Survey,” ACM Computing Surveys, vol. 52, no. 5, Sep. 2019.
11. M. Gygli, H. Grabner, and L. V. Gool, “Video summarization by learning submodular mixtures of objectives,” in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 3090–3098.
12. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, “Summary Transfer: Exemplar-Based Subset Selection for Video Summarization,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 1059–1067.
13. M. Ma, S. Mei, S. Wan, Z. Wang, D. D. Feng, and M. Bennamoun, “Similarity based block sparse subset selection for video summarization,” IEEE Trans. on Circuits and Systems for Video Technology, pp. 1–1, 2020.
14. Y. Li, T. Zhang, and D. Tretter, “An overview of video abstraction techniques,” Hewlett Packard, Technical Reports, 01 2001.
15. J. Calic, D. P. Gibson, and N. W. Campbell, “Efficient Layout of Comic-Like Video Summaries,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 17, no. 7, pp. 931–936, July 2007.
16. T. Wang, T. Mei, X. Hua, X. Liu, and H. Zhou, “Video Collage: A Novel Presentation of Video Sequence,” in 2007 IEEE Int. Conf. on Multimedia and Expo, July 2007, pp. 1479–1482.
17. C. Szegedy, Wei Liu, Yangqing Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich, “Going deeper with convolutions,” in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 1–9.
18. C. Szegedy, V. Vanhoucke, S. Ioffe, J. Shlens, and Z. Wojna, “Rethinking the inception architecture for computer vision,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 2818–2826.
19. A. Krizhevsky, I. Sutskever, and G. E. Hinton, “Imagenet classification with deep convolutional neural networks,” in Advances in Neural Information Processing Systems, 2012, pp. 1097–1105.
20. K. He, X. Zhang, S. Ren, and J. Sun, “Deep residual learning for image recognition,” in 2016 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2016, pp. 770–778.
21. K. Simonyan and A. Zisserman, “Very deep convolutional networks for large-scale image recognition,” in Int. Conf. on Learning Representations, 2015
22. K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, “Video Summarization with Long Short-Term Memory,” in Europ. Conf. on Computer Vision (ECCV) 2016, B. Leibe, J. Matas, N. Sebe, and M. Welling, Eds. Cham: Springer International Publishing, 2016, pp. 766–782
23. B. Zhao, X. Li, and X. Lu, “Hierarchical Recurrent Neural Network for Video Summarization,” in Proc. of the 2017 ACM on Multimedia Conf. (MM ’17). New York, NY, USA: ACM, 2017, pp. 863–871.
24. L. Lebron Casas and E. Koblents, “Video Summarization with LSTM and Deep Attention Models,” in MultiMedia Modeling, I. Kompatsiaris, B. Huet, V. Mezaris, C. Gurrin, W.-H. Cheng, and S. Vrochidis, Eds. Cham: Springer International Publishing, 2019, pp. 67–79.
25. J. Fajtl, H. S. Sokeh, V. Argyriou, D. Monekosso, and P. Remagnino, “Summarizing Videos with Attention,” in Asian Conf. on Computer Vision (ACCV) 2018 Workshops, G. Carneiro and S. You, Eds. Cham: Springer International Publishing, 2019, pp. 39–54.
26. Z. Ji, K. Xiong, Y. Pang, and X. Li, “Video Summarization With Attention-Based Encoder–Decoder Networks,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 30, no. 6, pp. 1709–1717, 2020.
27. Z. Ji, F. Jiao, Y. Pang, and L. Shao, “Deep attentive and semantic preserving video summarization,” Neurocomputing, vol. 405, pp. 200 – 207, 2020.
28. M. Rochan, L. Ye, and Y. Wang, “Video Summarization Using Fully Convolutional Sequence Networks,” in Europ. Conf. on Computer Vision (ECCV) 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 358–374.
29. L. Feng, Z. Li, Z. Kuang, and W. Zhang, “Extractive Video Summarizer with Memory Augmented Neural Networks,” in Proc. of the 26th ACM Int. Conf. on Multimedia (MM ’18). New York, NY, USA: ACM, 2018, pp. 976–983.
30. J. Wang, W. Wang, Z. Wang, L. Wang, D. Feng, and T. Tan, “Stacked Memory Network for Video Summarization,” in Proc. of the 27th ACM Int. Conf. on Multimedia (MM ’19). New York, NY, USA: ACM, 2019, p. 836–844.
31. M. Elfeki and A. Borji, “Video Summarization Via Actionness Ranking,” in IEEE Winter Conf. on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, Jan 2019, pp. 754–763.
32. B. Mahasseni, M. Lam, and S. Todorovic, “Unsupervised Video Summarization with Adversarial LSTM Networks,” in 2017 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), 2017, pp. 2982–2991.
33. E. Apostolidis, A. I. Metsai, E. Adamantidou, V. Mezaris, and I. Patras, “A stepwise, label-based approach for improving the adversarial training in unsupervised video summarization,” in Proc. of the 1st Int. Workshop on AI for Smart TV Content Production, Access and Delivery (AI4TV ’19). New York, NY, USA: ACM, 2019, pp. 17–25.
34. T. Fu, S. Tai, and H. Chen, “Attentive and Adversarial Learning for Video Summarization,” in IEEE Winter Conf. on Applications of Computer Vision (WACV), Waikoloa Village, HI, USA, January 7-11, 2019, 2019, pp. 1579–1587.
35. Y. Jung, D. Cho, D. Kim, S. Woo, and I. S. Kweon, “Discriminative feature learning for unsupervised video summarization,” in Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019.
36. L. Yuan, F. E. H. Tay, P. Li, L. Zhou, and J. Feng, “Cycle-SUM: CycleConsistent Adversarial LSTM Networks for Unsupervised Video Summarization,” in Proc. of the 2019 AAAI Conf. on Artificial Intelligence, 2019.
37. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, “Unsupervised Video Summarization via Attention-Driven Adversarial Learning,” in Proc. of the 26th Int. Conf. on Multimedia Modeling (MMM 2020). Cham: Springer International Publishing, 2020, pp. 492–504.
38. X. He, Y. Hua, T. Song, Z. Zhang, Z. Xue, R. Ma, N. Robertson, and H. Guan, “Unsupervised Video Summarization with Attentive Conditional Generative Adversarial Networks,” in Proc. of the 27th ACM Int. Conf. on Multimedia (MM ’19). New York, NY, USA: ACM, 2019, pp. 2296–2304.
39. M. Rochan and Y. Wang, “Video Summarization by Learning From Unpaired Data,” in 2019 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2019, pp. 7894–7903.
40. K. Zhou and Y. Qiao, “Deep Reinforcement Learning for Unsupervised Video Summarization with Diversity-Representativeness Reward,” in Proc. of the 2018 AAAI Conf. on Artificial Intelligence, 2018.
41. N. Gonuguntla, B. Mandal, N. Puhan et al., “Enhanced Deep Video Summarization Network,” in 2019 British Machine Vision Conf. (BMVC), 2019.
42. B. Zhao, X. Li, and X. Lu, “Property-constrained dual learning for video summarization,” IEEE Trans. on Neural Networks and Learning Systems, vol. 31, no. 10, pp. 3989–4000, 2020.
43. S. Cai, W. Zuo, L. S. Davis, and L. Zhang, “Weakly-Supervised Video Summarization Using Variational Encoder-Decoder and Web Prior,” in Europ. Conf. on Computer Vision (ECCV) 2018, V. Ferrari, M. Hebert, C. Sminchisescu, and Y. Weiss, Eds. Cham: Springer International Publishing, 2018, pp. 193–210.
44. Y. Chen, L. Tao, X. Wang, and T. Yamasaki, “Weakly Supervised Video Summarization by Hierarchical Reinforcement Learning,” in Proc. of the ACM Multimedia Asia (MMAsia ’19). New York, NY, USA: ACM, 2019.
45. K. Zhou, T. Xiang, and A. Cavallaro, “Video Summarisation by Classification with Deep Reinforcement Learning,” in 2018 British Machine Vision Conf. (BMVC), 2018.
46. Y. Yuan, T. Mei, P. Cui, and W. Zhu, “Video Summarization by Learning Deep Side Semantic Embedding,” IEEE Trans. on Circuits and Systems for Video Technology, vol. 29, no. 1, pp. 226–237, Jan 2019.
47. E. Apostolidis, E. Adamantidou, A. I. Metsai, V. Mezaris, and I. Patras, “AC-SUM-GAN: Connecting Actor-Critic and Generative Adversarial Networks for Unsupervised Video Summarization,” IEEE Trans. on Circuits and Systems for Video Technology, pp. 1–1, 2020.
48. Y. Jung, D. Cho, S. Woo, and I. S. Kweon, “Global-and-local relative position embedding for unsupervised video summarization,” in Europ. Conf. on Computer Vision (ECCV) 2020, A. Vedaldi, H. Bischof, T. Brox, and J.-M. Frahm, Eds. Cham: Springer International Publishing, 2020, pp. 167–183.
49. G. Yaliniz and N. Ikizler-Cinbis, “Using independently recurrent networks for reinforcement learning based unsupervised video summarization,” Multimedia Tools and Applications, vol. 80, no. 12, pp. 17 827–17 847, 2021. [Online]. Available: <https://doi.org/10.1007/s11042-020-10293-x>
50. P. Li, Q. Ye, L. Zhang, L. Yuan, X. Xu, and L. Shao, “Exploring global diverse attention via pairwise temporal relation for video summarization,” Pattern Recognition, vol. 111, p. 107677, 2021. [Online]. Available: <https://www.sciencedirect.com/science/article/pii/S0031320320304805>
51. B. Zhao, X. Li, and X. Lu, “TTH-RNN: Tensor-Train Hierarchical Recurrent Neural Network for Video Summarization,” IEEE Trans. on Industrial Electronics, vol. 68, no. 4, pp. 3629–3637, 2020.
52. S. Lal, S. Duggal, and I. Sreedevi, “Online video summarization: Predicting future to better summarize present,” in 2019 IEEE Winter Conf. on Applications of Computer Vision (WACV). IEEE, 2019, pp. 471–480.
53. W.-T. Chu and Y.-H. Liu, “Spatiotemporal Modeling and Label Distribution Learning for Video Summarization,” in 2019 IEEE 21st Int. Workshop on Multimedia Signal Processing (MMSP). IEEE, 2019, pp. 1–6.
54. Y. Zhang, M. Kampffmeyer, X. Zhao, and M. Tan, “DTR-GAN: Dilated Temporal Relational Adversarial Network for Video Summarization,” in Proc. of the ACM Turing Celebration Conf. (ACM TURC ’19) - China. New York, NY, USA: ACM, 2019, pp. 89:1–89:6.
55. M. Gygli, H. Grabner, H. Riemenschneider, and L. Van Gool, “Creating Summaries from User Videos,” in Europ. Conf. on Computer Vision (ECCV) 2014, D. Fleet, T. Pajdla, B. Schiele, and T. Tuytelaars, Eds. Cham: Springer International Publishing, 2014, pp. 505–520. [Online]. Available: <https://gyglim.github.io/me/>
56. Y. Song, J. Vallmitjana, A. Stent, and A. Jaimes, “TVSum: Summarizing web videos using titles,” in 2015 IEEE/CVF Conf. on Computer Vision and Pattern Recognition (CVPR), June 2015, pp. 5179–5187. [Online]. Available: <https://github.com/yalesong/tvsum>