IIETA International Information and Engineering Technology Association
Advancing the World of Information and Engineering

Revue d'Intelligence Artificielle
Vol. 35, No. 4, August, 2021, pp. 325-330
Journal homepage: http://iieta.org/journals/ria

Feature Extraction Model with Group-Based Classifier for Content Extraction from Video Data

Gowrisankar Kalakoti[1,2*], Prabakaran G[1]

[1] Department of Computer Science and Engineering, Annamalai University, Chidambaram 608002, Tamil Nadu, India
[2] Department of Information Technology, RVR & JC College of Engineering, Chowdavaram 522019, Guntur, Andhra Pradesh, India

Corresponding Author Email: gowrisankar508@gmail.com

https://doi.org/10.18280/ria.350407

Received: 12 November 2020
Accepted: 17 August 2021

Keywords:
content extraction, feature selection, group-based classifier, image extraction, video information, pixel classification

ABSTRACT

In today's PC illustration, numerous object locations of videos are quite critical duties to accomplish. Swiftly and reliably recognising and distinguishing the multiple aspects of a video is a crucial attribute for collaborating with one's condition (object). The core issue is that in theory, to ensure that no significant aspect is missing; all aspects of a content in a video must be scanned for elements on various different scales. It requires some investment and effort anyway, to really arrange the substance of a given content region and both time and computational limits that an operator can spend on classification are constrained. Two presumption procedures for accelerating the standard identifier are performed by the proposed method and demonstrate their capability by performing both identification efficiency and velocity. The main enhancement of our group-based classifier focuses on accelerating the grouping of **s**ub featur**es** by planning the problem as a selection procedure for consecutive features. The subsequent improvement gives better multiscale features to distinguish objects of all sizes without rescaling the information image from a video. Extracting contents from video is an assortment of successive images with a steady time interim. So video can give more data about contents in it when situations are changing regarding time. Along these lines, physically taking care of contents with features are very unimaginable. In the proposed work, it is suggested that a Group-based Video Content Extraction Classifier (GbCCE) extracts content from a video by extracting relevant features using a group-based classifier. The proposed method is distinct from conventional approaches and the findings indicate that better output is demonstrated by the proposed method.

1. INTRODUCTION

Video content recognition is a significant test activity that aims to recognise, interpret and extract pixels over a grouping of images called video within the image considered in Computer Vision. It helps with awareness, representing artifacts as opposed to human administrators testing PCs. It expects moving material to be contained in a video or exploration camera. It assists with comprehension, depict object as opposed to checking PC by human administrators. It expects to find moving Contents in a video or exploration camera. Content Extraction is the way toward finding content or various information utilizing a solitary camera, different cameras or given video document. Development of high quality of the imaging sensor, nature of the image and goals of the image are improved, and the exponential addition in calculation power is required to be made of new great calculation and its application utilizing object detection.

A difficulty of Object Detection and Tracking Objects in a general sense involves evaluating the area of a specific regions in progressive casings in a video arrangement. Appropriately distinguishing items can be an especially testing task, particularly since contents can have rather convoluted structures and may change based on size area and angle over resulting video outlines. Different calculations and plans have been presented in the couple of decades, that can follow problems in a specific video grouping, and every calculation has their own points of interest and disadvantages. Any content extraction calculation will contain errors which will in the end cause that extract irrelevant features of objects in video. The better calculations ought to have the option to limit this irregular feature extractions with the end goal that the tracker is precise over the time span of the application [1-4]. The video tracking process in frames is depicted in Figure 1.

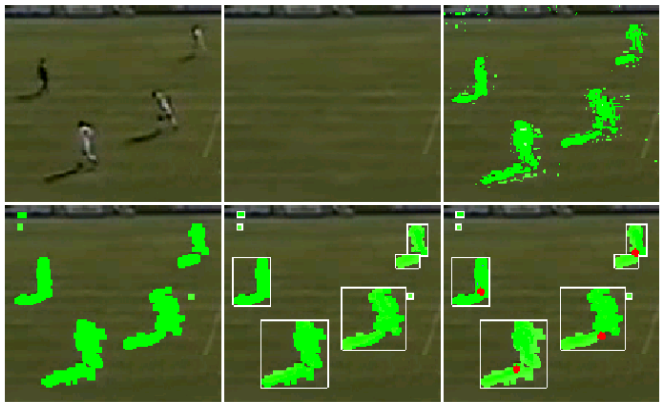

Figure 1. Video tracking

The system used for our particular observation is called visual consideration. A typical thought is that gatherings of neurons in different areas in the visual field go after limits. To figure out which area ought to be joined in, a specific arrangement of features should be processed in equal at each area. This pre-attentive stage is then trailed by further preparing of those areas that limits the opposition. There exist various speculations about which phases of handling are pre-attentive and which possibly happen when the improvements are joined in.

Feature Reduction is a method that can be applied to any classification issue. When managing a particular classification task earlier information can be used about the kind of information to accelerate arrangement [5, 6]. Two presumptions hold for most vision-based content identification methods:

(1) By far most of the investigated designs in an image has a place with the basic class and structure.
(2) The majority of the basic examples can be handily recognized from the items.

In view of these two presumptions, it is reasonable to apply a chain of importance of classifiers. Quick classifiers expel huge pieces of the foundation on the base and center degrees of the progression and a progressively exact method however slower classifier plays out the last location on the top level. This thought falls into the structure of content layout coordinating and identified with naturally stimulated fragment away at consideration based vision [7-10]. The content from the image is extracted and represented in Figure 2.

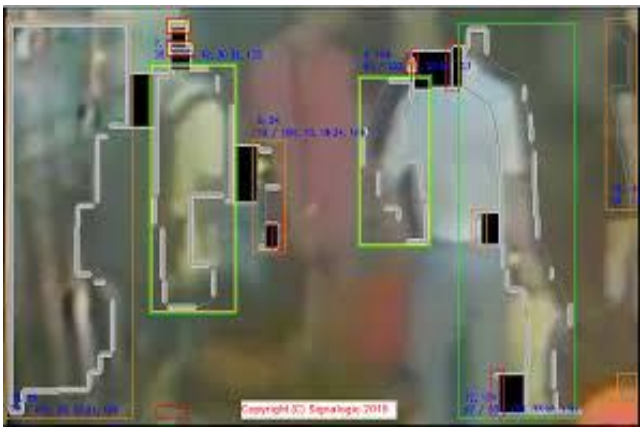

Figure 2. Content identification

Considering object identification and item acknowledgment with regards to AI based classification and feature selection, each frame in the video among all filtered frames could be considered as an information point. For object recognition and content extraction, in the preparation stage, a lot of images are gathered from images with marked content, which are encased by relevant object frames in which content frames and pixels that don't contain the items as negative models. Discernable features are then separated from the extracted frames to articulate to every model, and suitable classifiers are utilized on the extricated features to create a model. In the arrangement stage, the produced model is applied on a similar feature extraction method for every obscure content relevant frames to appoint its class name for accurate content extraction from video data.

2. LITERATURE SURVEY

For the instance of moving camera, it is significant that the strategy for moving item location considers not just all issues emerged in a fixed camera yet in addition certain challenges because of compensation of camera movement. This is the reason for a basic foundation with a trusting movement pay model can't effectively be applied for a moving camera. In fact, inaccuracy which is profoundly workable for a free development of the camera, causes the model displaying to come up with making a decent model for foundation and frontal area pixels. For recognizing content on images, one technique is to separate the developments brought about by moving items from those brought about by the camera. There are two fundamental classes of arrangement.

One depends on foundation displaying which attempts to make an appropriate foundation for each casing of the succession by utilizing a movement content extraction technique. Another is direction arrangement in which long directions are registered for features focuses utilizing a proper tracker and next a grouping approach is utilized to separate the directions having a place with similar content from those of basic models [11, 12]. Another procedure is to expand foundation of feature removal techniques dependent on low position and insufficient pattern deterioration created for the instance of static cameras for the instance of a moving camera. The foremost thought is that if certain coherency exists between a lots of image outlines, low position portrayal of the grid shaped by these edges contains this coherency and insufficient portrayal of this framework contains anomalies.

Since the content in video give changes, which are not quite the same as the foundation and can't be fitted into the low-position model of the foundation, they can be considered as anomalies for the low position portrayal. Subsequently, insufficient representation of the edges contains the moving Contents in these casings. In any case, it is dependent on the suspicion that the foundation is the equivalent for all casings, for example the camera is static. Despite the fact that, this strategy can't straightforwardly be applied for the instance of a moving camera, where the foundation changes between outlines, a change can be coordinated into the model so as to make up for the foundation movement brought about by the moving camera.

Khalil et al. [13**]** evaluated the works done under the general term of content location in video and ordered them as feature based, format based, classifier based and movement put together without any requirements with respect to camera movement. In the study, it has been acquainted different division strategies important with following contents in video and sorted item following into pixel extraction, content extraction and outline following and looked at the techniques in every classification. Yang et al. [14] separated item extraction strategies into form based, feature based and region based. Zhao et al. [15] additionally centered around following items by isolating it into three stages of content identification, object arrangement and item extraction and thought about the strategies proposed in every progression. In every one of these reviews, no limitations on camera movement were forced.

Celik and Bilge [16] proposes a face identification strategy dependent on instances of face images and non-face images. It gathered a lot of face formats and non-face layouts as the preparation set, utilizes a "Mahalanobis-like" metric as the component portrayal for each image, and utilizes a Multi-Layer Perceptron (MLP) model as the classifier to choose if

each checked frame is a face or non-face. It can identify countenances of various scales under different enlightenments. Be that as it may, it can just identify frontal appearances with high computational expense.

The region based identification procedures gives a profitable way to deal with and research development in an edge arrangement of video [17]. A casing area may be described as a lot of pixels having homogeneous qualities. It could be controlled by image division, which may be engaged around different object qualities like shading, edges, etc. Essentially, a region would be the image edge that is about by the projection of the object of venture onto the edge. Then again, a region could be the skipping box of the foreseen object under assessment.

Amelio et al. [18] recommends a calculation to segregate the moving Contents in video groupings and afterward introduced a standard based calculation. The fundamental trial results show the effectiveness of the calculation even in some confused circumstances, for example, new track, stopped track, track impact, and so forth. A content retrieval strategy without foundation extraction is examined. Since while separating content from video outline if there are little moving things in that outline they structure a mass in thresholding which make disarray in the event of features that mass as they are not of any utilization that can be diminished here. The author presents a video content extraction in PC vision, including structure prerequisites and an audit of strategies from video frame to following complicated, deformable items by learning models of shape and elements in an image.

3. PROPOSED METHOD

Content identification and Object detection are normally treated as two separate procedures. Content Detection in images depends on spatial appearance features while object detection in videos depends on both spatial appearance and ephemeral movement features. Noteworthy advancement has been made for object detection in 2D images utilizing CNN method [19, 20]. The standard thing "recognizing by innovation" for object detection necessitates that the Content is effectively recognized in the primary edge and every single resulting covering, and following is finished by "associate" recognition results. Performing Content Recognition through a private system is a very complex task [21].

The process of a content classification depends on upon the idea of the information that can be recognized from the photos. To perceive how image information is extracted, feature extraction is one of the basic advancements in the content identification process [22]. It authorizes to identify and extract data from an image quickly using a group-based classifier. The essential period of estimation requests over all scales and image regions. It can be completed by strategy for a distinction of Gaussian ability to perceive potentially relevant pixel points that are invariant to scale and direction and can be excluded. Relevant attributes identify with neighbor pixels of distinction of Gaussian channels at various scales [23]. Given a image portrayed as:

$$I(x,y,\lambda) = GP(x,y,\lambda) * I(x,y) + GDM \quad (1)$$

Here x,y are neighbor pixels and $\lambda$is the threshold in which pixels are extracted in the particular region, GDM is the Gaussian distribution mean used for relevant pixel extraction. The processing of the extracted pixels is arranged after performing preprocessing. The pixels extracted are compared with the neighbor pixels and noise values are removed using Gaussian distribution as:

$$G(x,y,\lambda) = \frac{1}{\pi\sigma^2} e^{-\frac{x^2+y^2}{\sigma^2}} + \mathrm{M}\,(x,y) \quad (2)$$

G is a Gaussian distribution variable, M is the mean of the extracted pixels whose result of convolving an image with a difference of Gaussian filter is given by:

$$GD(x,y) = L(x,y,k\sigma) + \frac{1}{\pi\sigma^2} M(x,y) + Wk(x,y) \quad (3)$$

The k neurons get M input parameters Xi and Yj. The neurons likewise has W weight parameters Wk (i,j). The weight parameters regularly incorporate a predisposition term that has a coordinating contribution with a fixed estimation of K. The data sources and loads are straightly consolidated and added as a group to form a cluster. The whole cluster is then taken care of to a classifier ψ that delivers the resultant neurons including hidden layers depicted in Figure 3.

$$C_k = \psi(s_k) = \psi\left(\sum_{j=0}^{m} w_{kj} a_j\right) + \sum_{k} p[x,y] \quad (4)$$

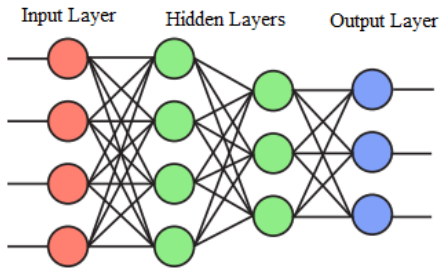

Figure 3. A fully-connected multi-layer neural network

ψ acts as quick detectors that are sensitive to certain types of pixels for example edges and contents in the images. Content edges are extracted and found across the visual field represented as a matrix represented in Figure 4.

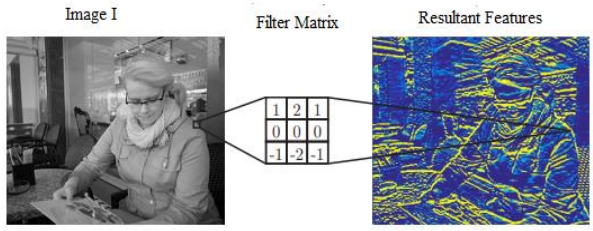

Figure 4. Horizontal edges using convolution filtering

The Discrete Convolution Operation (DCO) between an image I and a filter matrix g is defined as:

$$\begin{aligned} DCO[x,y] &= I[x,y] * g[x,y] \\ &= \sum_{n}\sum_{m} I[i,j]g[x-i,y-j] \\ &+ \sum_{k} W(p[x,y]) \end{aligned} \quad (5)$$

The convolution of a function Cf with another function Cp in a frame t is defined as:

$$\mathrm{Cf}(x * y)(t) = \int_{-\infty}^{\infty} Cp(x) * W(t-x)dx * dy$$
$$\mathrm{Cf}(\mathrm{I}(x * y)) = \sum_{x=-\infty}^{\infty} Cp(x) + (t-a) - W[x-i, y-j] \quad (6)$$

Convolution is regularly experienced with regards to image preparing, where x and y is the power of a given pixel and W is a 2-dimensional weighting capacity. t is normally non-zero just for a couple of qualities in the nearby pixels to the focal pixel of the edge and in this way the aggregate must be processed distinctly over those qualities rather than the entire image. If the weights W are for the most part positive and for instance an impact is accomplished by convolving the image with the portion where content is identified. On the off chance that again the pixel part's qualities are set comparative to,

$$\begin{pmatrix} -1 & 0 & 1 \\ -2 & 0 & 2 \\ -1 & 0 & 1 \end{pmatrix} or \begin{pmatrix} 1 & 2 & 1 \\ 0 & 0 & 0 \\ -1 & -2 & -1 \end{pmatrix} \quad (7)$$

In the wake of segmenting the video into pixels of different frames and removing key edges from each frame. The proposed quick classifier extracts the pixels from the edges quickly as only a segment of image where content is included is considered in the process. At that point we utilize key casings to partition the whole video into numerous "steady edge arrangements" by utilizing the ith key edge as the principal edge and utilizing the edge before I + 1th key edge as the last edge in the ith stable frame. The quantity of resultant stable edge arrangement in the video is equivalent to the absolute number of frames.

A video requires bigger feature set VFs i=1 FS(ai), At that point an absolute cluster is generated with the content as:

$$\mathrm{CS(VFs)} = \sum_{i=1}^{N} \alpha[P(x_i) + W(x)] + \sum_{i=2}^{N}(1-\alpha) + W(y) \quad (8)$$

For the first category of the proposed features performed using the quick classifier for content extraction, the variance of Images in a different color channel could be expressed as:

$$\mathrm{I(x,y)} = \frac{\sum_{i=1}^{M}\sum_{j=i}^{N}(P_{i,j,c=R} - \frac{\sum_{j=b}^{b+h}\sum_{i=a}^{a+w} P_{i,j,c=R}}{w \times h})^2}{(w \times h - 1)} + \mathrm{W(I(i))} - \left(\sum_{j=0}^{m} w_{kj}x_j\right) \quad (9)$$

The content loss function is minimized during the extraction part with consideration of only pixels that are relevant. The loss function is represented as:

$$\mathrm{CLR(I(x,y))} = \frac{1}{N}\sum_{i=1}^{N} \log p(a_i, b_i) + \lambda \frac{||w||_1}{||w||_2} + Wk(x,y) \quad (10)$$

As the loss function is less, the content is extracted from the images accurately and in less time as the image is segmented into parts. The segmented images contain relevant content and the proposed classifier quickly extracts the features of the content. The classifier assigns weights for the clusters and the clusters having high priority has been extracted and formed as a content cluster.

4. RESULTS

The proposed Group-based Classifier for Content Extraction from Video is implemented in Python**,** the dataset used is available in the URL https://open-video.org/. The proposed method is compared with the traditional Content Based Image Retrieval (CBIR) method and the results show that the proposed method is exhibiting better results than the traditional methods.

The proposed strategy accomplishes noteworthy improvement in creating precise recommendations for inflexible and non-unbending items. For content extraction from video, which commonly utilizes transient objects can be figured to improve the nature of the content recommendations. At the end consolidating the extracted content in video can accomplish higher accuracy with a similar number of approvals from video images or a similar review rate with the less number of proposition. The video is divided into frames for extracting content from the image. The video frames segmentation is depicted in Figure 5.

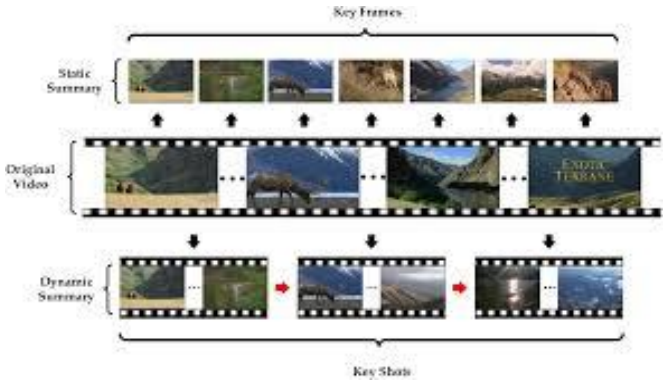

Figure 5. Frame segmentation

The features from the video frames are extracted and formed as a cluster. The process is depicted in Figure 6.

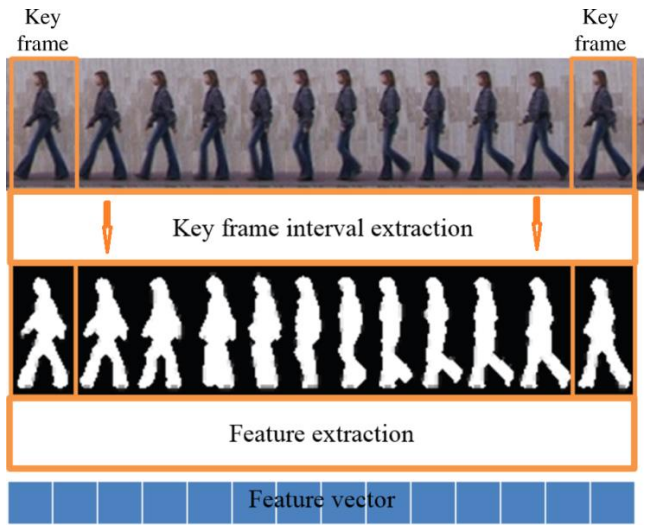

Figure 6. Feature extraction

4.1 Precision

In the Proposed video retrieval precision is characterized by calculation with online and offline video retrieval process. The group of retrieved through the quantity of relevant videos are called as the precision.

$$\text{Precision}(\text{Relevant} \,||\, \text{Retrieved}) = \frac{\text{relevant(video)} \cap \text{retrieved(video)}}{\text{retrieved(video)}} \quad (11)$$

The performance evaluation with reduction method of the proposed method is depicted in Table 1.

Table 1. Performance evaluation with reduction method

Reduction Method	Metric	Precision	Recall
Performed	Euclidean	0.6	0.76
Not Performed	Euclidean	0.4	0.62

The number of features extracted from the video to identify the content from image is accurate so the content identification from the video is quick and accurate. The feature extraction accuracy is depicted in Figure 7.

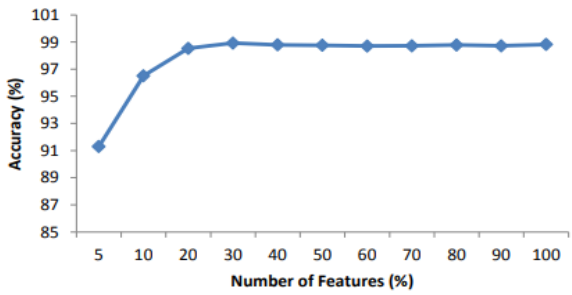

Figure 7. Precision accuracy as number of features is varied

The edges of the content in the image frame of a video are extracted by detecting edges of the image and the irrelevant features are excluded. The process is depicted in Figure 8.

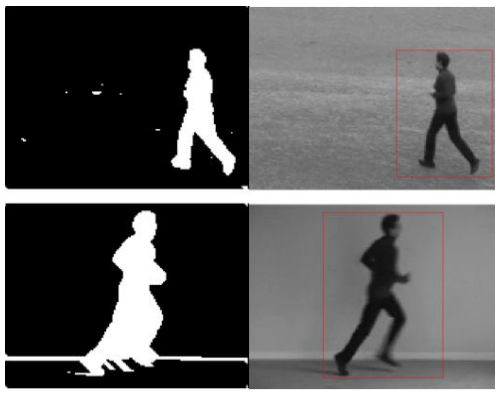

Figure 8. Content extraction

The overall performance evaluation of the proposed method is depicted in Table 2. The performance of the proposed method is better when contrasted with traditional methods.

The accuracy level of the proposed method is compared with the traditional method and the results show that the proposed method is more accurate than the traditional methods. The accuracy levels are depicted in Figure 9.

Table 2. Overall performance evaluation

Video Frames	Metric	Precision	Recall	F-Measure
100	Euclidean	0.52	0.7	0.75
200	Euclidean	0.56	0.84	0.68
300	Euclidean	0.58	0.86	0.88
400	Euclidean	0.62	1	0.89
500	Euclidean	0.83	0.85	0.82

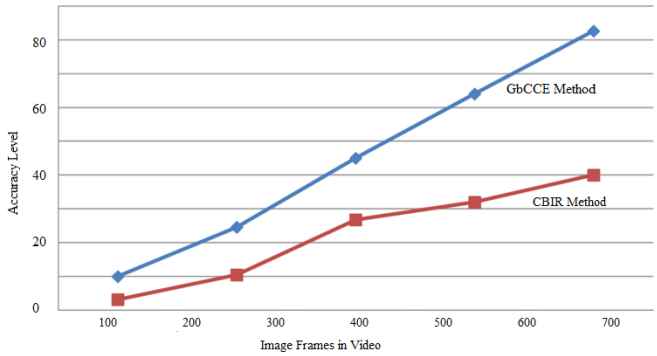

Figure 9. Accuracy levels

5. CONCLUSION

Content identification is a significant errand in computer vision field. Content extraction in video image is accomplished by extracting relevant features. Content Extraction is the way toward finding content or various information utilizing a solitary camera, different cameras or given video document. Development of high quality of the imaging sensor, nature of the image and goals of the image are improved. Recognizing and distinguishing the different features from a video quickly and dependably is a significant expertise for collaborating with one's condition. The fundamental issue is that in principle, all features of a content in a video must be scanned for objects on numerous different scales to ensure that no important feature is missed. The proposed Group-based Classifier for Content Extraction from Video exhibits better performance in extraction of relevant features and content extraction from the video. The proposed method exhibits 96% of accuracy in accurate content extraction and in future the proposed method can be extended by directly considering video input and identifying content wherever required. The number of features can be reduced and the neurons generated can be increased for better outcomes.

REFERENCES

[1] Markowska-Kaczmar, U., Kwaśnicka, H. (2018). Deep learning—A new era in bridging the semantic gap. In: Kwaśnicka H., Jain L. (eds) Bridging the Semantic Gap in Image and Video Analysis. Intelligent Systems Reference Library, vol 145. Springer, Cham. https://doi.org/10.1007/978-3-319-73891-8_7

[2] Zhou, W.G., Li, H.Q., Tian, Q. (2017). Recent advance in content-based image retrieval: A literature survey. https://arxiv.org/abs/1706.06064

[3] Riaz, F., Jabbar, S., Sajid, M., Ahmad, M., Naseer, K., Ali, N. (2018). A collision avoidance scheme for autonomous vehicles inspired by human social norms. Computers & Electrical Engineering, 69: 690-704.

https://doi.org/10.1016/j.compeleceng.2018.02.011
[4] Amelio, A. (2019). A new axiomatic methodology for the image similarity. Applied Soft Computing, 81: 105474. https://doi.org/10.1016/j.asoc.2019.04.043
[5] Zhang, C., Cheng, J., Tian, Q. (2019). Unsupervised and semi-supervised image classification with weak semantic consistency. IEEE Transactions on Multimedia, 21(10): 2482-2491. https://doi.org/10.1109/TMM.2019.2903628
[6] Shi, X., Sapkota, M., Xing, F., Liu, F., Cui, L., Yang, L. (2018). Pairwise based deep ranking hashing for histopathology image classification and retrieval. Pattern Recognition, 81: 14-22. https://doi.org/10.1016/j.patcog.2018.03.015
[7] Alzu'bi, A., Amira, A., Ramzan, N. (2017). Content-based image retrieval with compact deep convolutional features. Neurocomputing, 249: 95-105. https://doi.org/10.1016/j.neucom.2017.03.072
[8] Kondylidis, N., Tzelepi, M., Tefas, A. (2018). Exploiting tf-idf in deep convolutional neural networks for content based image retrieval. Multimedia Tools and Applications, 77(20): 30729-30748. https://doi.org/10.1007/s11042-018-6212-1
[9] Bushra, Z., Rehan, A., Nouman, A., Mudassar, A., Sohail, J., Kashif, N., Awais, A., Gwanggil, J. (2018). Intelligent image classification-based on spatial weighted histograms of concentric circles. Computer Science and Information Systems, 15(3): 615-633. https://doi.org/10.2298/CSIS180105025Z
[10] Yang, H.F., Lin, K., Chen, C.S. (2018). Supervised learning of semantics-preserving hash via deep convolutional neural networks. IEEE Transactions on Pattern Analysis and Machine Intelligence, 40(2): 437-451. https://doi.org/10.1109/TPAMI.2017.2666812
[11] Kalakoti, G., G, P. (2020). Key-frame detection and video retrieval based on DC coefficient-based cosine orthogonality and multivariate statistical tests. Traitement du Signal, 37(5): 773-784. https://doi.org/10.18280/ts.370509
[12] Zhu, L., Shen, J., Xie, L., Cheng, Z. (2017). Unsupervised visual hashing with semantic assistant for content-based image retrieval. IEEE Transactions on Knowledge and Data Engineering, 29(2): 472-486. https://doi.org/10.1109/TKDE.2016.2562624
[13] Khalil, T., Akram, M.U., Raja, H., Jameel, A., Basit, I. (2018). Detection of glaucoma using cup to disc ratio from spectral domain optical coherence tomography images. IEEE Access, 6: 4560-4576. https://doi.org/10.1109/ACCESS.2018.2791427
[14] Yang, S., Li, L., Wang, S., Zhang, W., Huang, Q., Tian, Q. (2019). SkeletonNet: A hybrid network with a skeleton-embedding process for multi-view image representation learning. IEEE Transactions on Multimedia, 21(11): 2916-2929. https://doi.org/10.1109/TMM.2019.2912735
[15] Zhao, W., Yan, L., Zhang, Y. (2018). Geometric-constrained multi-view image matching method based on semi-global optimization. Geo-Spatial Information Science, 21(2): 115-126. https://doi.org/10.1080/10095020.2018.1441754
[16] Celik, C., Bilge, H.S. (2017). Content based image retrieval with sparse representations and local feature descriptors: A comparative study. Pattern Recognition, 68: 1-13. http://dx.doi.org/10.1016/j.patcog.2017.03.006
[17] Nie, X., Jing, W., Cui, C., Zhang, J., Zhu, L., Yin, Y. (2019). Joint multi-view hashing for large-scale near-duplicate video retrieval. IEEE Transactions on Knowledge and Data Engineering, 32(10): 1951-1965. https://doi.org/10.1109/TKDE.2019.2913383
[18] Amelio, L., Janković, R., Amelio, A. (2018). A new dissimilarity measure for clustering with application to dermoscopic images. Proceedings of the 2018 9th International Conference on Information, Intelligence, Systems and Applications (IISA), IEEE, Zakynthos, Greece, pp. 1-8. http://doi.org/10.1109/IISA.2018.8633672
[19] Zhang, C., Cheng, J., Tian, Q. (2017). Structured weak semantic space construction for visual categorization. IEEE Transactions on Neural Networks and Learning Systems, 29(8): 3442-3451. https://doi.org/10.1109/TNNLS.2017.2728060
[20] Zhang, C., Cheng, J., Tian, Q. (2018). Semantically modeling of object and context for categorization. IEEE Transactions on Neural Networks and Learning Systems, 30(4): 1013-1024. https://doi.org/10.1109/TNNLS.2018.2856096
[21] Qi, G., Wang, H., Haner, M., Weng, C., Chen, S., Zhu, Z. (2019). Convolutional neural network based detection and judgement of environmental obstacle in vehicle operation. CAAI Transactions on Intelligence Technology, 4(2): 80-91. http://doi.org/10.1049/trit.2018.1045
[22] Maddumala, V., Arunkumar, R. (2020). Big data-driven feature extraction and clustering based on statistical methods. Traitement du Signal, 37(3): 387-394. https://doi.org/10.18280/ts.370305
[23] Lindeberg, T. (2013). Scale selection properties of generalized scale-space interest point detectors. Journal of Mathematical Imaging and Vision, 46: 177-210. https://doi.org/10.1007/s10851-012-0378-3