ISSN: 0067-2904

Automatic Image and Video Tagging Survey

Suha Dh. Athab*, Abdulamir Abdullah Karim
Department of Computer Science, University of Technology, Baghdad, Iraq,

Received: 30/3/2022 Accepted: 4/12/2022 Published: 30/9/2023

Abstract

Marking content with descriptive terms that depict the image content is called "tagging," which is a well-known method to organize content for future navigation, filtering, or searching**.** Manually tagging video or image content is a time-consuming and expensive process. Accordingly, the tags supplied by humans are often noisy, incomplete, subjective, and inadequate. Automatic Image Tagging can spontaneously assign semantic keywords according to the visual information of images, thereby allowing images to be retrieved, organized, and managed by tag. This paper presents a survey and analysis of the state-of-the-art approaches for the automatic tagging of video and image data. The analysis in this paper covered the publications on tagging in Scopus and the Web of Science databases from 2008 to 2022.

Keywords Social media, Image tagging, Video tagging.

دراسة لوضع العلامات على الصور والفيديو بصورة تلقائية

سها ظاهر عذاب *، عبد الامير عبد الله عبد الكريم
قسم علوم الحاسوب، الجامعة التكنلوجية، بغداد، العراق

الخلاصة

يمكن تمييز الصور والفديو بمصطلحات وصفية لتصوير محتواها. عملية اضافة كلمات وصفية لمحتوى البيانات هي عملية معروفة لأغراض تنظيم المحتوى, التصفح ,البحث والتصفية . يعد وضع علامات على الصورة او الفديو يدويا عملية مكلفة وتستغرق الكثير من الوقت علاوة على كونها غير منتظمة وتفتقر الى توحيد المصطلحات. في حين ان عملية اضافة العلامات بصورة تلقائية من خلال تعيين كلمات رئيسية ذات دلالة لمحتوى الصورة تسهم وبشكل كبير وفعال في استرداد الصور وتنظيمها وادارتها بصورة سريعة واكثر دقة. تقدم هذه الورقة البحثية مسحا

وتحليلا لأحدث الاساليب الخاصة بوضع العلامات بصورة تلقائية على الصور والفديو. غطى تحليل هذه الورقة البحثية مجموعة من الابحاث العلمية للفترة من 2008 الى 2022 التي تم نشرها ضمن سكوباس وشبكة قواعد البيانات العالمية.

1. Introduction

Image tagging involves analyzing the objects inside the image and assigning a tag that can properly depict the image content. On the other hand, video tagging is the process of adding a tag to each keyframe in the video [1, 2]. Image tagging makes internet searching easier, additionally enables the quick organization of a tremendous number of images, and makes them

*Email: cs.20.09@grad.uotechnology.edu.iq

easily accessible. Due to the significant expansion of multimedia content already available and continuously uploaded and shared on social media platforms, machine learning algorithms have a significant role in making such information easier to find and link [3]. This review consists of definitions of the goals of each paper, the datasets, and the techniques used by different researchers to improve image and video tagging efficiency. Additionally, the results of each paper have been illustrated. The organization for the rest of this paper was organized as follows: Section 2 presents the most common datasets used in the image and tag domains as well as the well-known evaluation metrics used. In Section 3, the summarized tables review the significant common methods in video and image tagging. Section four has the discussion part. The survey conclusion is presented in Section 5.

2. Datasets & Evaluation metrics

2.1 Datasets

Multiple researchers use tagging in various domains, such as developing an automatic attendance system for college students [4], elderly activities in the K-Log center for Alzheimer's patients' video tagging [5], and movie segmentation [6]. For this kind of research, a special dataset was collected to fulfill the main objective of the research [7]. For tagging as a fundamental objective, the researchers used the following datasets: Corel5K, NUS-WIDE, YOLOv3, YFCC100M, ESP Game, IAPRTC-12.5, Tencent Advertisement Video, Chicago Face Database (CFD), and Event. Table 1 describes the datasets in detail.

Table 1: Tagging datasets

Datasets name	No. of images/videos	Description
***Corel5K* [1]**	5000 image	The average manual tag assigned for each image is three and a half keywords from 260 predetermined terms.
NUS-WIDE [8]	269648 images	The National University of Singapore created a real-world web image dataset. The dataset details are shown below: (1) 269,648 images joined with 5,018 tags. (2) Low-level features include a colored histogram, an edge direction histogram, wavelet texture, block-wise color moments, and a bag of words based on SIFT descriptions extracted from the images. (3) For evaluation purposes, a ground truth of 81 concepts was supplied.
YFCC100M [9]	nearly100million	The dataset, which contains parts of Flickr images combined with hashtags and GPS coordinates
Tencent Advertisement Video[10]	10000 videos	500 videos for training, the videos labeled using timestamps, and 500 videos for a test. The average length of the videos is 42.74 ± 14.16 seconds. A series of multiple tags for each scene, which represent the classes each scene belongs to. There is no overlapping between scenes.
IAPRTC-12.5*[11, 12]*	19627 images	The captions associated with each image are used to infer the tag. The dataset has the following classes: sports, actions, people, animals, cities, landscapes, and many other aspects.
ESP *game[11, 12]*	20770 images	Diverse types of images exist, such as logos, drawings, and personal photos. A total of 268 tags are included in the dataset.
MediaQ and GeoVid projects[13]	2397 videos	The statistics of a dataset can be summarized as follows: The dataset contains 208,978 video frames with an average length per video of 72.14 sec.
UCF101[9, 14]	13320 videos	The dataset has 101 classes and 5 categories (human-object interaction, body-motion, human interaction, playing musical instruments, and sports).

Approximately 25 research papers were analyzed from 2008 to 2022 within this research**.** All the analyzed papers were within the domain of image and video tagging. The most frequent datasets used were Corel5 and Nus-wide due to the diversity of the classes in each dataset. In addition, each dataset is provided with a manually tagged label, which eases the training process of the tagging system. Moreover, a valuable set of low-level features was extracted and provided with the datasets to increase the accuracy of the tagging process.

2.2 Evaluation metrics

Most researchers have recently used per-image metrics, such as precision (Eq. 1), recall (Eq. 2), F1-measure (Eq. 3), accuracy (Eq. 4), and mean average precision (mAP) (Eq. 5), to accurately evaluate tagging performance [12, 15]. The metric values are averaged over all the images in the test dataset to obtain average per-image metrics [14]. The definitions of per-image metrics are as follows:

$$precision = \frac{TP}{TP+FP} \tag{1}$$

$$Recall \quad = \frac{Tp}{Tp + Fp} \tag{2}$$

$$F_1 \ = 2 \times \frac{Precision \times Recall}{Precision + Recall} \tag{3}$$

$$Accuracy \quad = \frac{Tp + TN}{Tp + Fp + TN + FN} \tag{4}$$

$$mAP = \frac{1}{N} \sum_{k=1}^{N} AP_k \tag{5}$$

Where TP is the number of tags that are predicted by the model and match correctly. TN is the number of tags that are not predicted by the model and are not part of the ground truth. FP denotes the number of tags predicted by the model but not included in the ground truth, whereas FN doesn't predict the tag, but it is part of the ground truth. F1-measure combines Precision and recall. AP_k average $precision$ of class. k the number of classes

3. Literature review

Many researchers use diverse methods and a tremendous number of features and techniques for image and video tagging purposes. Borth, Damian, et al. (2008) use the six Tamura features. These are contrast, directionality, coarseness, line-likeness, regularity, and roughness for automatic video tagging [16]. One of the interesting methods was the implicit tagging that was introduced by J. Jiao and M. Pantic in 2010 (which is the technique to tag multimedia data based on a user's nonverbal reactions, such as facial expressions and head gestures). Nineteen facial points were employed for tracking facial expressions and used to judge explicit tagging [17]. Other researchers, such as Yang et al. (2011), transfer knowledge between images and videos, with tags assigned by using structures embedded within both the image and video spaces [15]. Binti Zakaria in 2012 used City Landscape Identifier (CLI) to represent image content by exploiting the edge direction and then developed a classifier that can be used for automatically tagging images with "buildings" or "non-buildings" tags. Gomez, Raul, et al. (2020) used a large set of images, tags, and geographical coordinates, to redesign a model used for tagging and retrieving images when the query combined the hashtag and location information, to reduce the effort and work time for the caregivers (CGs) for logging and monitoring Alzheimer's Disease (AD) patients. A multi-modal fusion based on machine-assisted human tagging of

videos and the object detection model was introduced by Lee, Chanwoong, et al. in 2020. [5] Prediction and segmenting a movie at the viewer's choice. [6] proposed a real-time attendance system using image tagging to overcome the wastage of time in queues for biometrics or face-scanning using the LBPH algorithm. [4] le 2, and 3 Describe in detail 23 research papers published in Scopus and Web of Science databases from 2008 to 2022 for image and video data tagging, respectively.

Table 2: image tagging research

Author	Aim	Dataset	Method	Result
Jiao and M. Pantic 2010 [17]	Use spontaneous nonverbal reaction to convey the correctness of tags associated with images	28 images	19 points on the face were extracted and tacked to get two types of features (appearance and geometric) that are used to represent facial expression.	60% accuracy
Leong, C. W and et.al 2010[18]	Explore the use of several natural language resources to construct an image tagging model	300 image-text pairs were collected from the web	Presented three extractive approaches for the task of image tagging 1.Wikipedia Salience uses a graph-based method to tag the image with the keywords extracted from a text. 2. Flickr Picturality Given the text (T) of an image, use it to retrieve the most frequent associated Flickr tags using getRelatedTags API with a given word. 3. Topical Modeling the Pachinko Allocation Model (PAM) was used to model the topics in a text. Use the PAM model to infer a list of *super-topics* and *sub-topics*	accuracy 92%
Yang, Yang, et al. 2011[8]	Improving the performance of image tagging.	NUS-WIDE	A near-duplicate clustering algorithm was adopted for tag aggregation. A weighted association algorithm is used to infer correlations between tags. Near duplicates of the image were retrieved to generate its candidate tags and the initial corpus relevance score from a test image. Multi-tag association rules are used to get relevant tags.	accuracy 62.59
Binti Zakaria, 2012[19]	developed a classifier that can be used for automatically annotating images with a "buildings" or "non-buildings" tag.	The training set from Flickr contains 210 images (105 buildings and 105 nonbuildings). Two test sets, each containing 534 buildings and 506 nonbuilding images	A Bayes-based machine learning tool from Microsoft called Infer.NET was used. Known building and non-building images are submitted to low-level feature extractors for color and line features. The inference engine uses the training data in order to generate prediction values in the range 0 to 1.0 for each image in the set. The City Landscape Identifier analyzer identifies edges in 72 directions. Indoor/outdoor classification is done by inferring the LUV color space. City or non-city is classified by observing the edge direction histogram. Sunset, forest, and mountain classifications are	accuracy 91.67%.

			identified by using color features in HSV space.	
Chen, et al 2013[12]	Given the training images annotated with incomplete tags, the goal is to learn and infer the full list of tags	1. *Corel5K* 2. ESP *gam* IAPRTC	For each dataset, fifteen distinct visuals and descriptors (one Gist descriptor, six global color histograms+, and eight local bag-of-visual-words features) were extracted.	**Corel5K** precision 32% ,recall 43%, F1 37%, **ESPgame** precision 46%, recall 22%, F1 30% **IAPR** Precision 47%, recall 26% , F1 34%
Gomez, Raul, et al 2020 [9]	Designing image tagging and retrieving model when the query combined the hashtag and location information,	24.8M from the YFCC100M dataset separated into a validation set of 250K and a test set of 500K. Images have an average of 4.25 hashtags.	Initially, learn image representations using hashtags by training the CNN model. For the training, three procedures were used: first, multi-label classification; after that, softmax multi-class classification; and, subsequently, hashtag embedding regression. Finally, train multimodal models to score triplets of these three modalities.	Accuracy 44.00.
Patil, Vishal, et al 2020[20]	Proposed a real-time attendance system using image tagging to overcome the wastage of time in queues for biometrics or face scanning,	A face dataset of college students is created at this point, in which 45 pictures are taken of each scholar in the category	**Preprocessing:** consists of two phases: initially, face recognition using the Haar classifier and the LBP cascade classifier face detection algorithm, then resizing the face to a constant pixel. **Training phase**: applying LBPH to train and build a model using the real-time dataset of images. Histograms for every image are created here. The LBPH algorithm was chosen to process the data correctly in real time. The essential idea is to divide the LBP picture into local regions and extract a histogram for each.	Accuracy 75%
	Combining visual and annotated information to achieve a better tagging performance	1.Corel, 2. NUS-WIDE 3. Flickr	To produce accurate image tagging, the aggregated network and cooperative training were integrated. The visual information was exploited for the tagging approach; the features were acquired by a modified neural network method. The correlated information for the tagging words is then fully utilized.	**precision** for Corel 5K, NUS−WIDE, and Flickr was 0,45, 0.5, 0.64 respectively. **recall** Corel 5K, NUS−WIDE, and Flickr were 0.57, 0.6, and 0.66 respectively. **F1** for Corel 5K, NUS−WIDE, and Flickr were 0.5, 0.55, 0.65

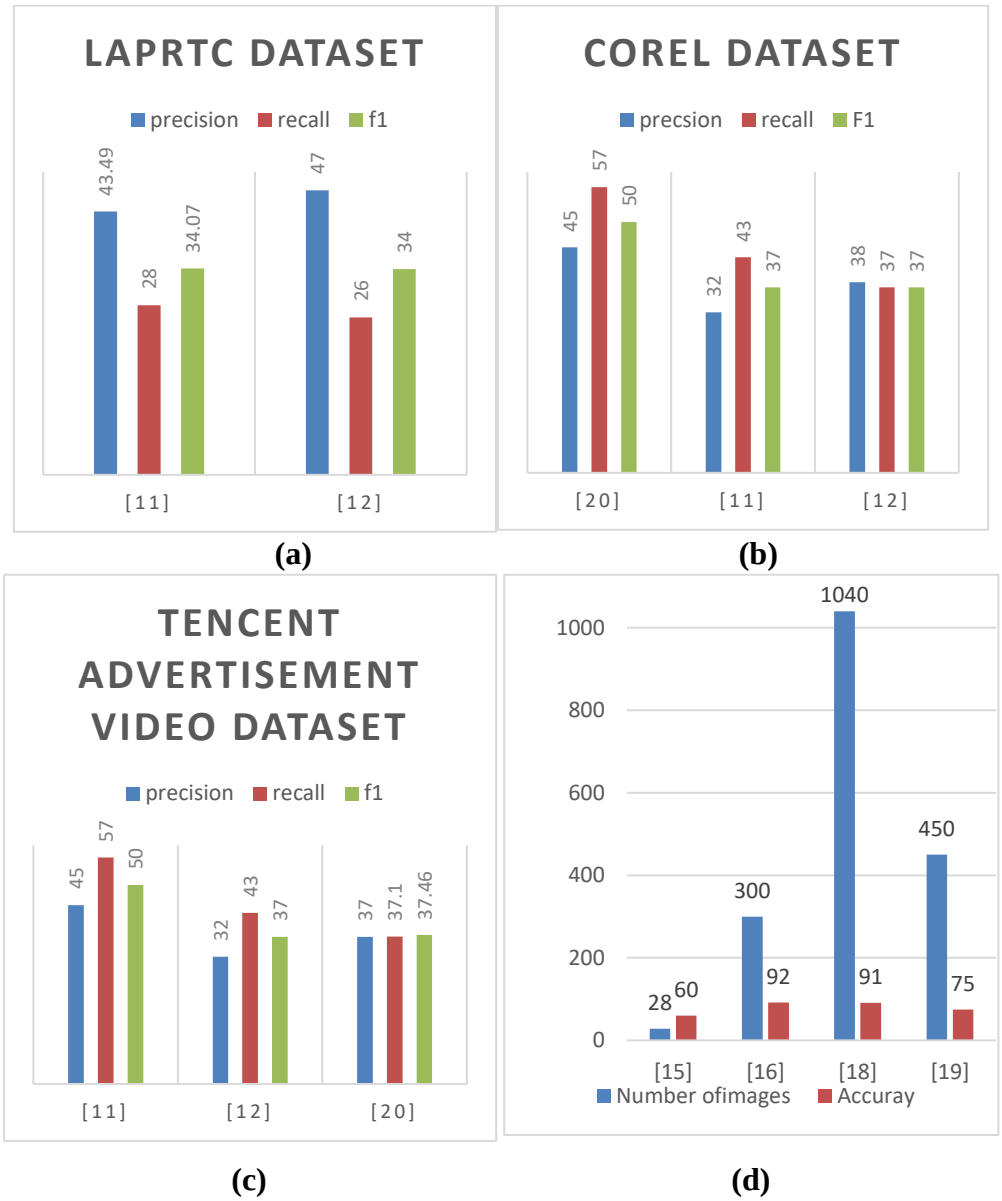
Table 3: video tagging

Author	Aim	Dataset	Method	Result
Borth, Damian, et al.2008 [16]	Automatic tagging of video	2200 videos with a total duration of 194 hours were downloaded from YouTube	The shot boundary detection technique and an intra-shot clustering of frames were combined in addition to a different combination of color histograms and Tamura features.	Mean Average Precision (mAP) 34.2%
Siersdorfer, S., San Pedro, J., & Sanderson, M. 2009[7]	Exploit a high redundancy (duplicated) content to get relationship information of duplicated videos	38, 283 videos contain over 2900 hours of video.	automatically tag videos in folksonomies by merging content-based copy retrieval and tag propagation methods.	Clustering with k-Means accuracy 55%. Classifying using linear support vector machines (SVMs) accuracy of 76%
Yang, Yang, et al 2011[15]	Exploiting the underlying structures embedded within both image and video spaces used to infer a tag	1. NUSWIDE 2. MIRFlickr and two video datasets, 1. Kodak 2. TRECVID	Transfer tag knowledge between the image and video used. The cross-media video tagging scheme proposed the following steps to be implemented: a. Bridging Image and Video b. Cross Media Tagging. C. Laplacian Matrix and Multiple Kernels	mAP 77.6%
Yao, Ting, et al. 2013[21]	Use the click-through data to mine the video relationship for tagging purposes.	test samples randomly selected from 2,500 URLs.	For characterizing video similarity and annotating videos, the co-click (footprint for user behavior) and polynomial semantic similarity, with two tagging methods, were used.	Precision 0.419 Recall 0.808 F1 0.504
To, Hien, et al 2016[13]	Create an approach to search and filter big multimedia data, specifically geo-tagged mobile videos, for context-aware AR applications.	Media and GeoVid projects	Three approaches are used for incorporating video content into Augmented Reality applications: pre-defined, on-demand, and suggested content by hotspots.	Accuracy 90%
Wu, Shan, Shangfei Wang, and Zhen Gao 2017[22]	Capture the inherent dependencies among video content, which are crucial for personalized video emotion tagging.	CP-QAE-I database	The model learns the relationships, characteristics, and tag behaviors of video among the content during the training. Then the emotion tag of the video is generated by the model after it has learned meaningful topics.	Accuracy 72.2% FI score 64.3
Wang, Shanghai, Shiyu Chen, and Qiang Ji. 2017[23]	develop a method for video emotion tagging in which three different feature spaces can be obtained from training samples, and only one is required for testing samples	1. Lirisaccede 2. Ustcervs 3. Mahnobhci 4. Deap	combine similarity constraints on the emotion classifiers from videos and the emotion classifiers from available physiological signals to capture the nature of the relationships among users' physiological responses, video content, and emotion labels.	Deep dataset F1-score 0.606, accuracy 0.658. Users dataset: F1-score 0.745, Accuracy 0.868 Mahnobhci dataset: F1-score 0.470, Accuracy 0.548 Lirisaccede dataset: F1-score 0.751, accuracy 0.757
Yang, Wenmian, et al 2017[24]	propose an unsupervised video tag extraction algorithm for online video tags using Time-sync comments	A total of 227,780 random comments related to music, sports, and movie about 120000 used for a training set.	Semantic similarities and timestamps were used for generating the semantic association graph (SAG) of the time-sync comments. Then it clusters the comments into sub-graphs of different topics and assigns weight to each comment based on SAG.	F-score 0.4104 mAP 0.3518

Ilyas, S., & Rehman, H. U. 2019 [14]	A method for indexing and retrieving videos	UCF101 dataset	Capturing video content by exploiting keyframe information The keyframes used to train the parameters of the convolution NN The tag is infrared according to the trained parameters.	Accuracy 99.8% F1-score 96.2
Patwardhan, Abhishek, et al 2019[25]	A new approach to automatic video tagging	103 videos for 13 domains from YouTube	It involves video segmentation to extract a representative frame, and then the tag extracted from the segment is used to predict semantic similarity. The inferred tag is finally used to annotate the input video.	precision 65.51%
Takeda, Hiroshi, Soh Yoshida, and Mitsuji Muneyasu 2019[26]	Improve performance of retrieving video based on tag	YouTube-8M dataset	The tag reference score is calculated by considering the tag frequency of occurrence for a tag, and then a tag neighbor voting algorithm is used.	The author did not mention numerical results and described them (as "effective and efficient")
Lee, Chanwoong, et al2020[5]	To reduce the effort and the ime of work for the caregivers (CGs) for taking care of and logging the activities of Alzheimer's Disease patients	K-Log Centre surveillance videos	The YOLO-v3 object detection is integrated with HAR models, which are used for automatic tagging of the surveillance videos.	An accuracy of 81.4% for live video.
Khan, Umair Ali, et al 2020[6]	Segmenting a movie at the viewer's choice.	A tag vocabulary contains 50 movie tags. Then a dataset was created for each tag. From several movies, around 700 images for each tag were collected.	A deep learning-based technique for predicting the most relevant tags for a movie and segmenting the movie concerning the predicted tags Inception-V3 was used to train a pre-trained CNN for task transfer learning. Subsequently, a frame detection algorithm was used.	mAP 76.50% F1-score 0.7551.
Suzuki, Tomoyuki, and Antonio 2021[10].	Create a pipeline for segmenting and tagging videos	Tencent Advertisement Video	A bi-level approach that initially provides the boundaries of the scenes and then merges a confidence score for each segmented scene. The predicted tag of the class proposed for segmented video ads	mAP 0.86%

As mentioned earlier, the researchers used different global datasets and various accuracy measures in line with the nature and idea of the research; some of the researchers created their own custom datasets. To compare them and know the best method, the comparison will not be fair; however, we can draw approximate conclusions about the best method by comparing the researchers who signed up for the same dataset and scale as shown in Figures (1-a), (1-b), and (1-c), where the researchers used the datasets Corel, LAPRTC, and ESP game, respectively.

While the researchers who have created the custom dataset are working, their work will be compared based on the size of the dataset that has been created, as shown in Figure 1–d:



Figures 1(a, b, c, d): side by side comparison of different researchers to Corel, LAPRTC, and Tencent Advertisement video datasets respectively, Figure(1-d) side by side comparisons for the custom-created datasets

4. Discussion

There are diverse methods used by researchers for image and video tagging. Most of the methods used a supervised approach, as in [4] [5] [7] [12] [14] [16] [19] [20], whereas few researchers adopted an unsupervised approach, as in [6] [8] [15] [25]. Both the supervised and unsupervised approaches used for image and video tagging leveraged different types of features. Some researchers, such as [16, 19], adopted low-level features followed by a simple classification method to map between image and tag. Exploiting the low-level features used to convey the correctness of the tag associated with the images is better than a random guess. The low-level features have bad noise resistance. To create classifiers for generating a wide range of tags, it is necessary to use more powerful low-level features, such as visual terms. Colored histogram features are used for tagging purposes; sometimes the histogram features are

combined with LBP and Haar to enhance the results, as in [12, 20]. The methods are computationally efficient in that they require only one matrix inversion per iteration. However, the prediction loss for each tag is weighed equally, which leads to the overall loss being dominated by contributions from more frequent tags, sacrificing the prediction accuracy of rare tags. The main challenge in the fully automated HAR scenarios is collecting the datasets, which demands precise human and object detection during the testing phase. A real-time method used by [5, 20] shows that the existing strategies do not operate properly in instances of distinctive illumination. The methods that gave the best result used CNN as the main algorithm for mapping between the image features and the semantic tag, as in [9, 27] [14] [6]. It is pertinent to mention that the data used for training CNN-based models is manually tagged. In light of the spread of epidemics and infectious diseases, the future scope of tagging domains within the field of robots in sanitary isolation rooms is to help patients and reduce infection rates for medical staff.

5. Conclusion

Marking the image with descriptive terms is also called "tagging." A huge range of digital enterprises depend on photo tagging to manage their visual assets; e-commerce, stock photo databases, booking and travel platforms, traditional and social media, and a variety of other businesses require adequate and efficient image sorting systems. Additionally, tagging is useful to individuals; personal photo libraries can be difficult to organize and search through without user-friendly image categorization and tagging. Decades ago, traditional indexing was performed by a librarian. An intriguing alternative to traditional indexing methods was collaborative tagging, or "folksonomy," which is the practice of allowing users to attach tags to data; however, collaborative tagging suffers from slowness, expense, being highly subjective, and the inability to scale to multi-million image libraries; thus, there is a strong interest among computer vision researchers in the development of robust and efficient automatic image and video tagging systems. A discriminative model (nearest neighbor), a generative model, or a deep learning model could be used for automatic tagging. A discriminative model describes tagging as a multi-label classification problem. Each label trains a separate classifier using the features extracted from the image. Later, the trained classifier predicts a tag for the test image. By learning joint distributions over visual and contextual features, generative models accurately detect dependency between visual features and associated tags. The generative model produced a remarkable contribution to the development of tagging; the complexity of the generative model's algorithms was the reason for its inability to achieve optimization in tag prediction. The Nearest Neighbor models, in contrast to the generative models, were motivated to be widely used in the tagging domain due to their simplicity. The nearest neighbor focused on selecting similar neighbors and then assigning the tags to the test image. Image-to-image, image-to-tag, or both similarities could be used. Subsequently, a greedy label transfer mechanism is employed to assign the tag, which is the method of selecting the tags according to the co-occurrence and frequency factors of the nearest neighbors. Many computer vision tasks showed high-quality overall performance by adopting deep learning-based methods that extracted effective feature vectors from images to make perfect mappings between the image and the semantic tag.

References

[1] J. Chen, P. Ying, X. Fu, X. Luo, H. Guan, and K. J. I. T. o. M. Wei, "Automatic tagging by leveraging visual and annotated features in social media," *IEEE Transactions on Multimedia*, vol. 24, pp. 2218-2229, Jan 2021.

[2] M. M. J. I. J. o. S. Mijwil, "Implementation of Machine Learning Techniques for the Classification of Lung X-Ray Images Used to Detect COVID-19 in Humans," *Iraqi Journal of Science*, vol. 62, pp. 2099-2109, Jul. 2021.

[3] S. D. Athab, and N. H. J. I. J. o. S. Selman, "Localization of the Optic Disc in Retinal Fundus Image using Appearance Based Method andVasculature Convergence," *Iraqi Journal of Science*, vol. 61, no. 1, pp. 164–170, Jan. 2020.

[4] T. Sutabri, A. K. Pamungkur, R. E. J. I. J. o. M. L. Saragih, and Computing, "Automatic attendance system for university student using face recognition based on deep learning," *International Journal of Machine Learning and Computing*, vol. 9, no. 5, pp. 668-674, 2019.

[5] C. Lee, H. Choi, S. Muralidharan, H. Ko, B. Yoo and G. J. Kim, "Machine Assisted Video Tagging of Elderly Activities in K-Log Centre," *2020 IEEE International Conference on Multisensor Fusion and Integration for Intelligent Systems (MFI)*, 2020, pp. 237-242, doi: 10.1109/MFI49285.2020.9235269.

[6] U. A. Khan et al., "Movie Tags Prediction and Segmentation Using Deep Learning," *in IEEE Access*, vol. 8, *pp*. 6071-6086, *2020, doi: 10.1109/ACCESS.2019.2963535.*

[7] S. Siersdorfer, J. San Pedro, and M. Sanderson, "Automatic video tagging using content redundancy," *In Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*, pp. 395-402.

[8] Y. Yang, Z. Huang, H. T. Shen, and X. J. W. W. W. Zhou, "Mining multi-tag association for image tagging," *Springer*, vol. 14, no. 2, pp. 133-156, 2011.

[9] R. Gomez, J. Gibert, L. Gomez, and D. Karatzas, "Location sensitive image retrieval and tagging," *Computer Vision – ECCV 2020: 16th European Conference, Glasgow, UK*,Proceedings, Part XVI, pp. 649–665, Aug 2020. https://doi.org/10.1007/978-3-030-58517-4_38

[10] T. Suzuki, and A. J. a. p. a. Tejero-de-Pablos, "Video Ads Content Structuring by Combining Scene Confidence Prediction and Tagging," *arXiv*:2108.09215 , 2021.

[11] W. Li, H. Song, H. Zhang, H. Li, and P. Wang, "The Image Annotation Refinement in Embedding Feature Space based on Mutual Information," *International Journal of Circuits, Systems and Signal Processing*, vol. 16, pp. 191-201, 01/08, 2022.

[12] M. Chen, A. Zheng, and K. Weinberger, "Fast image tagging," *In International conference on machine learning*, PMLR, pp. 1274-1282, 2013.

[13] H. To, H. Park, S. H. Kim and C. Shahabi, "Incorporating Geo-Tagged Mobile Videos into Context-Aware Augmented Reality Applications," *2016 IEEE Second International Conference on Multimedia Big Data (BigMM)*, pp. 295-302, 2016. doi: 10.1109/BigMM.2016.64.

[14] S. Ilyas, and H. U. Rehman, "A deep learning based approach for precise video tagging," *2019 15th International Conference on Emerging Technologies (ICET)*, IEEE, pp. 1-6, 2019.

[15] Y. Yang, Y. Yang, Z. Huang, and H. T. Shen, "Transfer tagging from image to video," *In Proceedings of the 19th ACM international conference on Multimedia*, pp. 1137-1140, 2011.

[16] D. Borth, A. Ulges, C. Schulze, and T. M. Breuel, "Keyframe Extraction for Video Tagging & Summarization," *In Informatiktage*, pp. 45-48, 2008.

[17] J. Jiao, and M. Pantic, "Implicit image tagging via facial information," *In Proceedings of the 2nd international workshop on Social signal processing 2010*,pp. 59-64, Oct. 2010.

[18] C. W. Leong, R. Mihalcea, and S. Hassan, "Text mining for automatic image tagging," *In Coling 2010: Posters*, pp. 647-655, 2010.

[19] L. Q. binti Zakaria, P. Lewis, and W. Hall, "Automatic image tagging by using image content analysis," *International Conference on Informatics and Applications 2012*, pp. 91-101, 2012.

[20] V. Patil, K. Kapadia, A. Khokrale, and P. Jain, "Intelligent College Attendance System Using Image Tagging," *In Proceedings of the 3rd International Conference on Advances in Science & Technology (ICAST)*, 2020.

[21] T. Yao, T. Mei, C.-W. Ngo, and S. Li, "Annotation for free: Video tagging by mining user search behavior," *Proceedings of the 21st ACM international conference on Multimedia*, pp. 977-986, 2013.

[22] S. Wu, S. Wang, and Z. Gao, "Personalized video emotion tagging through a topic model," *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, IEEE, pp. 2866-2870, 2017.

[23] S. Wang, S. Chen, and Q. J. I. T. o. A. C. Ji, "Content-based video emotion tagging augmented by users' multiple physiological responses," *IEEE Transactions on Affective Computing*, vol. 10, no. 2, pp. 155-166, 2017.

[24] W. Yang, N. Ruan, W. Gao, K. Wang, W. Ran, and W. Jia, "Crowdsourced time-sync video tagging using semantic association graph," *2017 IEEE International Conference on Multimedia and Expo (ICME)*, IEEE, pp. 547-552, 2017.

[25] A. A. Patwardhan, S. Das, S. Varshney, M. S. Desarkar, and D. P. Dogra, "ViTag: Automatic video tagging using segmentation and conceptual inference," *In 2019 IEEE Fifth International Conference on Multimedia Big Data (BigMM)*, pp. 271-276, Sept. 2019.

[26] H. Takeda, S. Yoshida, and M. Muneyasu, "Tag-based Video Retrieval with Social Tag Relevance Learning," *In 2019 IEEE 8th Global Conference on Consumer Electronics (GCCE)*, pp. 869-870, Oct. 2019.

[27] J. Chen, P. Ying, X. Fu, X. Luo, H. Guan and K. Wei, "Automatic Tagging by Leveraging Visual and Annotated Features in Social Media," *in IEEE Transactions on Multimedia*, vol. 24, pp. 2218-2229, 2022, doi: 10.1109/TMM.2021.3055037.