Stage art, Movie making (Сценічне мистецтво, кіно)

UDC: 791.43:004.93

Neryanov Pavlo

Professional Videographer

Route One Group Company

(East Dundee, IL, USA)

ORCID: 0009-0003-4525-8048

METHODS OF STRUCTURING VISUAL NARRATIVE IN SHORT VIDEO CONTENT

Summary. *The article analyzes the specifics of structuring visual narrative in short video content, in particular through the use of modern technologies to build a logical and emotional sequence of frames. It is established that the key methodological approach to structuring visual narrative in short video content is the formation of a narrative sequence that provides the logic of the development of the visualized story. It is found that the traditional linearity of chronological presentation can be supplemented or violated in order to achieve a cognitive or emotional effect, and visual forms, in particular infographics, require alternative means of organizing the narrative due to the lack of a clear timeline. It is determined that the use of deep learning technologies, in particular recurrent neural networks (RNN) and long short-term memory networks (LSTM), allows forming a coherent visual narrative by identifying logical connections between frames. The effectiveness of multimodal models that combine text and image analysis for identifying key plot elements is proven. The prospects of using graph knowledge models for representing narrative as a network of concepts, themes, and metadata are substantiated, which ensures flexibility and adaptability of the structure of the visual narrative. Four types*

of narrative structures of a short video are distinguished: a chain of facts, a chain of hypotheses, a chain of events, and a chain of arguments. Six binary attributes are characterized that reflect the structural and perceptual features of a visual narrative, including completion, plot structure, type of presentation, temporal structure, perspective of vision, and level of interactivity. It is summarized that the use of structural-typological classification contributes to a deeper analysis of the mechanisms of constructing a visual narrative and increases the effectiveness of video communication within short-length content. The methods of clustering and generalization of features used to form a coherent visual narrative are systematized. It is found that traditional clustering algorithms (K-Means, K-Median) have limitations in processing complexly structured data, while agglomerative clustering and density-based methods (DBSCAN) provide higher adaptability to the topology of the input information. The feasibility of using modern models of generalization of features, in particular tree-shaped decision-making algorithms, convolutional neural networks (CNN), models with attention mechanisms and visual transformers, which provide semantic depth and integrity of the visual narrative, is substantiated.

Key words: *visual narrative, short-length video content, structuring methods.*

Problem statement. In today's digital environment, video content plays a leading role in communication and cultural production, driven by the rapid development of streaming platforms, social networks, and mobile technologies. Among the variety of video formats, short videos attract particular attention as an effective tool for conveying ideas or emotions in a limited amount of time. Given the fast pace of modern life and changes in how audiences perceive information, the question of how to effectively construct a narrative in a short video format is becoming increasingly relevant.

Despite the growing interest in the analysis of visual narratives, the question of methods for structuring them in short video content remains under-researched in scientific terms. Most existing approaches are based on classical narrative models, which are not always effective in the context of short time and high video dynamics, necessitating the development of new methods adapted to the specifics of short video, taking into account the nature of digital content, multimodality, and interactivity of modern video formats. In addition, existing studies rarely combine narrative, cognitive, and technological aspects of visual content structuring, which limits their applicability in real-world video production. Therefore, there is a need for a comprehensive study of effective methods for constructing visual narratives that can meet the demands of modern audience perception and utilize the latest technological capabilities.

In this regard, structuring visual narrative in short-form video content requires the use of specific methods that ensure an organic combination of content, emotional intensity, and visual expressiveness. Therefore, it is necessary to conduct a thorough analysis of methods for structuring visual narratives in short-form video content in order to solve the problem of optimizing structural methods for constructing visual narratives.

Analysis of recent studies and publications. The issue of structuring visual narratives in video content is attracting growing interest among researchers in the fields of cognitive science, computer science, and visual communication. A significant contribution to the theoretical foundation of visual narratives has been made by Koun N. [1], which examines the cognitive model of the structure of visual storytelling, in particular the sequence of frames, their syntactic organization, and the role of transitions between visual units. The author substantiates the existence of a kind of "grammar" of visual narrative that functions independently of the verbal context.

Aspects of the plot composition of visual messages are revealed in the work of Phillips D. [2], who, based on the concept of eight basic plots, analyzes the principles of storytelling in the natural sciences, emphasizing the universality of plot models in the transmission of complex information. Parallels between cinematic and cartographic narratives can be traced in the studies of Mühlengauz I. [3] and Moknik F.-B., Ferber D. [4], where the map is considered as a means of visual storytelling with its own structure, genre features, and means of aesthetic influence.

The informational and visual aspects of narrative structuring are analyzed in the works of Gershon N., Page W. [5] and Gallmann D. et al. [6], who prove the importance of the sequence of visual elements, the selection of compositional accents, and the logic of data presentation in creating a convincing visual message. In particular, Gallmann D. proposes a typology of narrative sequences that allows formalizing the structure of video content from the viewer's perspective.

The generalization of approaches to cartographic storytelling presented in the work of Roth R.E. [7] is also valuable, where the genres and tropes of map-centric narratives are classified, and the limits of the application of such approaches in a broader multimedia and information context are outlined. Also important for our research are the works of Singh A., Sharma D.K. [8], which analyze methods for generalizing image collections, and Chen D., Zhuge H. [9], which propose approaches to multimodal summarization of documents that include both textual and visual information.

The work of Lotfi F. et al. [11] provides a systematic overview of methods and tools for image-based visual storytelling, which forms the conceptual basis for the development of short videos focused on structuring content through sequential visual fragments. Equally important is the work of Cao R. et al. [12], which investigates the use of narrative structures in data videos, in particular the structural principles of data integration, imagery, and emotional coloring, which allows for the creation of a

highly informative and engaging storyline. Despite the abundance of research highlighting individual aspects of visual narrative, the question of a comprehensive methodology for structuring narrative in short-form video content, which combines conciseness, aesthetics, and cognitive accessibility, remains understudied, which in turn determines the relevance of our research.

Aim of the study. The aim of this article is to analyze contemporary methods of structuring visual narratives in short video content, taking into account cognitive, compositional, and informational and communicational aspects, as well as to identify effective approaches to constructing a coherent, consistent, and aesthetically expressive video sequence within the constraints of limited time.

Presentation of the main material. One of the key methodological approaches to structuring visual narrative in short video content is the formation of a narrative sequence that determines the logic of the unfolding of the visualized story. In scientific literature, narrative is interpreted as the ordering of key elements characteristic of most plots [1]. In the field of cinema, narrative sequence is usually formed on the basis of the chronological order of events, which together form a storyline [2]. This approach determines the linearity of the presentation, which is one-dimensional in nature, although narrative authors often resort to disrupting chronology in order to achieve a distinct emotional or cognitive effect, in particular through the use of flashbacks or foreshadowing [3].

On the other hand, visual forms of information presentation, in particular cartographic or infographic materials based on spatial metaphors, are characterized by two-dimensionality, which results in the absence of a clear time axis as a basic tool for constructing a storyline [4]. Under such conditions, there is a need for a conscious and balanced selection of narrative components that will ensure the linearity and consistency of the perception of the visual story. That is why video content producers must actively use visual structuring tools that can compensate for

the absence of traditional chronological logic, thus forming a coherent and understandable narrative message [5-7].

One important aspect of structuring visual narratives in video content is the ability of technologies to detect and interpret temporal relationships between frames that form a sequence of events. Unlike static images, video contains dynamic information that is realized through temporal linearity. This allows deep learning architectures, such as recurrent neural networks (RNNs) and long short-term memory (LSTM) networks, to be used to detect logical connections between visual fragments, which helps create a coherent visual narrative [8]. In the context of visual storytelling, recurrent neural networks (RNN) and long short-term memory (LSTM) networks, thanks to their ability to store information about previous states, enable the sequential detection of logical and narrative connections within a video sequence. This allows not only to analyze the dynamics of an image, but also to form a coherent storyline, which is critical for building a coherent visual narrative..

In the context of short videos, such approaches allow not only to automate the selection of key moments, but also to maintain plot integrity, ensuring logical sequence and meaningful connection of the visual story. Thus, taking into account the sequence of events in time is an important factor in forming a clear and logically structured narrative in video content.

When structuring a visual narrative, it is important to consider the multimodal nature of modern short-form video content, which often combines images, text, video, and hyperlinks. Chen et al. [9] proposed a method for summarizing multimodal content that combines text and image analysis using hierarchical recurrent neural networks (RNN) and convolutional neural networks (CNN). This model allows for the simultaneous processing of verbal and visual data, which facilitates the accurate detection of key plot elements in a narrative.

One promising direction for structuring visual narratives involves the use of graph knowledge models, which integrate semantic information from text and images in short videos. In such a model, the narrative is presented as a network of interconnected elements—concepts, objects, themes, captions, and metadata—that together form a structured story. This allows individual storylines to be identified and the narrative to be adapted to the viewer's needs or interpretations [10]. The use of graph analysis algorithms makes it possible to view the entire network as a single semantic field within which individual subgraphs or paths function as basic plot units. This approach enables the automated construction or reconstruction of stories tailored to a given query. Ultimately, the graph knowledge model serves as the basis for the formation of a personalized multimodal visual narrative adapted to the specifics of short-form video content [11].

In modern short-form video content, particularly in data-driven videos, the narrative structure is increasingly moving away from traditional artistic models (plot-setting) and instead forming around the logic of information presentation. This approach requires the adaptation of classic storytelling ideas and the creation of new types of structures that better suit the functional and communicative tasks of visual content. It is worth highlighting four main types of narrative structures that are characteristic of short visual stories:

- Chain of facts – a sequential presentation of empirical data or facts that support the key message of the video.
- Chain of hypotheses – an ordered presentation of hypothetical assumptions that model possible situations or scenarios.
- Chain of events – constructing a video around a sequence of actual or reconstructed events.
- Argument chain – a presentation of logically justified positions, which may include both arguments and counterarguments.

Usually, videos are based on a single narrative structure, although it is possible to combine elements from other types. This approach to structuring allows for greater persuasiveness, clarity, and effectiveness of short-form video content, especially when dealing with abstract or complex topics.

When analyzing short video content, it is useful to identify a number of narrative attributes that allow for a systematic classification of narrative models (see Fig. 1). The proposed typology is based on the identification of six key binary categories that reflect the structural and functional characteristics of visual narrative. Each of these attributes represents a separate aspect of narrative structuring and provides an analysis of both the content and formal levels of its construction. The classification covers key dimensions of narrative organization: completion, plot structure, type of presentation, temporal structure, perspective, and level of interactivity, which demonstrates the interdependence between the compositional and content components of the narrative, determining how the viewer perceives the video material.

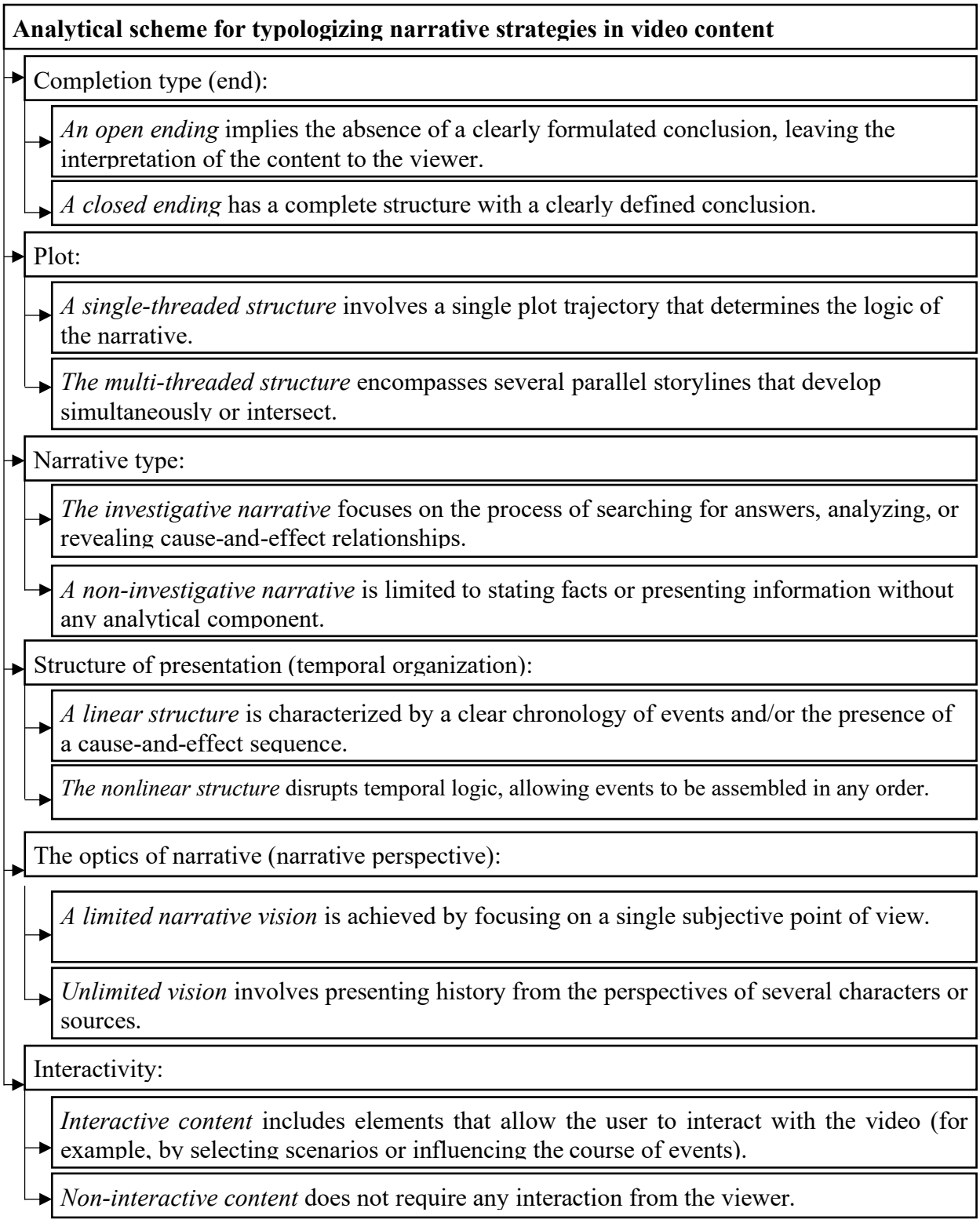

Fig. 1. Analytical diagram of the typology of narrative strategies in video content

Notes: compiled by the author based on source: [12]

Figure 1 allows us to systematize the main parameters of narrative construction in video content, providing a scientific understanding of both the structural and perceptual characteristics of audiovisual communication. The systematization of these attributes contributes to a more accurate understanding of the mechanisms of visual narrative construction and expands the possibilities for its further analysis and practical application in the field of short-form video production.

In the context of short-format visual storytelling, the narrative structure is formed on the basis of a number of means (see Table 1) that ensure effective visualization and transmission of content to the viewer. The main structuring tools include symbolic visual elements, reenactment of events, real visual cues, voice-over, annotations, musical accompaniment, and attention-grabbing devices.

The use of visual and narrative means in the structure of short video content indicates a steady trend towards the integration of multimodal tools that provide not only aesthetic content but also cognitive effectiveness of communication. In a scientific context, these tools are factors in the formation of the semiotic field of content, where the combination of visual, audio, and text components serves to convey knowledge, ideas, or emotions with a high degree of audience engagement. From the perspective of contemporary media linguistics and visual rhetoric, these tools are mechanisms for emphasizing, modeling attention, and forming interactive interaction between the media text and the audience. Their systematic organization within a video product contributes to the creation of a multifunctional narrative space where image, sound, text, and movement do not compete but complement each other. At the same time, the effectiveness of such means depends largely on the integrity of the conceptual design and the director's strategy.

Table 1

Main types of visual and audiovisual means in the structure of video narrative

A means of visual narrative	Characteristics	Functional purpose
Symbolic visual elements	Virtual modeling of objects using 2D graphics, 3D models, or VR environments.	Facilitates compact and clear presentation of complex information through conditional visual reproduction.
Reconstruction of events	Visualization of a chronological or reconstructed sequence of events in 2D, 3D, or VR formats.	Provides dynamism and immerses the viewer in the action through the use of camera movement and spatial scenes.
Real visual cues	Use of real-world photo or video clips.	Increases the reliability and authenticity of visual content, particularly in documentary or educational formats.
Voice-over	Audio commentary accompanying images provides explanations or interpretations.	Helps organize perception, enhances emotional coloring, and structures the visual sequence.
Abstract	Graphic or text inserts that explain events or visual components.	They perform a cognitive function – they facilitate the interpretation of complex or specialized information.
Music	Music that matches what's happening in the video.	It performs the role of emotional modulation, creates the appropriate atmosphere, and heightens dramatic tension.
Attention focusing tools	Include close-ups, zooming, color contrasts, animation, or dynamic movement.	They provide visual emphasis on key elements and help focus the viewer's attention.

Notes: compiled by the author based on source: [12]

The comprehensive use of visual narrative tools in short videos shows how we're moving from a linear to a multi-structured way of presenting info. These approaches fit the needs of today's viewers, who are used to being visually literate and ready to take in content in different ways.

The process of constructing visual narratives is a complex, multi-component task that involves the classification, analysis, and integration of both visual and textual data. Table 2 systematizes modern approaches to clustering and generalizing features that are key elements in the formation of a coherent visual narrative. Clustering methods, in particular the K-Means and K-Median algorithms, are aimed at identifying groups of similar objects, but demonstrate limited effectiveness in cases of complex or uneven data structures. In such cases, it is advisable to use more flexible algorithms, in particular agglomerative clustering and density-based methods (DBSCAN), which are capable of adapting to data topology.

Feature generalization, on the other hand, is achieved using models capable of forming highly informative data representations. Such models include tree-based decision-making algorithms (such as Random Forest), general-purpose artificial neural networks, convolutional neural networks (CNN), which work effectively with spatial image structures, as well as models with an attention mechanism, that allow computational resources to be automatically focused on the most relevant fragments of the input signal. Visual transformers deserve special attention, as they integrate attention to all elements of an image, providing a deep semantic understanding of the content and supporting the formation of a coherent, structured narrative.

Table 2

Methods of clustering and generalizing features for constructing visual narratives

Method	Description	Application
Clustering methods		
K-Means	An algorithm that groups elements by their average values, forming cluster centers.	Effective for processing large volumes of data with uniform distribution, allows identifying similar elements in visual fragments.

K-Median	An algorithm similar to K-Means, but based on median values, which reduces sensitivity to outliers that can distort clustering results.	Used to structure data in the presence of anomalies or uneven distributions.
Hierarchical clustering	A method of constructing a tree-like structure by merging or splitting clusters.	Useful for analyzing complex visual data with a multi-level hierarchical nature.
Agglomerative clustering	It starts with individual elements, gradually combining them into larger clusters.	Used to form a structure of relationships between objects, taking into account local similarities.
Dividitional clustering	Derived from a whole array of data, which is divided into subgroups.	This method is useful when it is necessary to decompose complex sets of visual objects.
Probabilistic clustering	Algorithms that allow an element to belong to several clusters with a certain probability.	The algorithm is effective for modeling fuzzy boundaries between clusters in visual content.
Density-based methods (e.g., DBSCAN)	Focused on identifying areas of high density in data.	Allows detecting clusters of arbitrary shape, especially in cases of uneven distribution of visual features.
Methods of generalizing characteristics		
Tree-like decision-making models	Building a hierarchy of decisions based on selected characteristics.	Used to classify scenes or objects based on certain characteristics.
Hidden Markov Models (HMM)	Statistical models for sequential processes with probabilistic transitions between states.	Used to analyze frame dynamics and sequential changes in video content.
Artificial neural networks	Multilayer structures that perform learning based on input data.	Used to generate image descriptors and build semantic representations.
Methods for measuring similarity	Algorithms that use different metrics to determine the similarity between elements.	Enable effective comparison of visual fragments or frames for grouping purposes.

CNN models (Convolutional Neural Networks)	Convolutional neural networks that work effectively with spatial image structures.	Widely used for object recognition, scene classification, and video frame vectorization.
Attention-based methods	Architectures that focus computation on the most relevant elements of the input data.	Improve scene analysis accuracy by allowing key elements of the frame to be highlighted.
Vision Transformers	Models based on the attention mechanism and using complete image information.	Ensure seamless data integration, effective for generating consistent

Notes: compiled by the author based on source: [13-31]

In general, the effective combination of clustering and feature generalization methods is key to creating structured, informative, and conceptually complete visual stories that enable deeper analysis of complex multimodal data.

Conclusions. Based on the conducted research, it can be concluded that structuring visual narrative in short video content is a key factor in ensuring the integrity and comprehensibility of the storyline in conditions of limited time. It has been established that the effectiveness of narrative sequencing depends on the consistency of spatial and temporal logic, as well as on the level of integration of visual and verbal components. It has been found that modern technologies, in particular recurrent neural networks (RNN, LSTM) and multimodal models, significantly improve the accuracy of identifying logical connections between frames and contribute to the synchronization of narrative elements. It has been established that the typology of narrative structures, in particular chains of facts, events, hypotheses, and arguments, allows for a more accurate classification of visual stories. It has been found that the use of binary attributes in the structural analysis of video sequences contributes to the development of adaptive information presentation strategies. Thus, the formation of an effective visual narrative requires an

interdisciplinary approach that combines technological tools, narrative models, and the communicative goals of video content.

The practical significance of this study lies in the fact that the conclusions and recommendations formulated by the author can be effectively used to improve methods for constructing visual narratives in short-form video content and to develop more adaptive script structures for digital media. Further research in this area should focus on an in-depth study of the mechanisms of interaction between visual and verbal components in short-form video content, taking into account the cognitive characteristics of perception.

References

1. Cohn, N. (2013). Visual narrative structure. *Cognitive Science, 37*(3), 413–452. https://doi.org/10.1111/cogs.12016.

2. Phillips, J. (2012). Storytelling in Earth sciences: The eight basic plots. *Earth-Science Reviews, 115*(3), 153–162. https://doi.org/10.1016/j.earscirev.2012.09.005.

3. Muehlenhaus, I. (2014). Looking at the big picture: Adapting film theory to examine map form, meaning, and aesthetic. *Cartographic Perspectives, (77)*, 46–66. https://doi.org/10.14714/cp77.1239.

4. Mocnik, F.-B., & Fairbairn, D. (2018). Maps telling stories? *The Cartographic Journal, 55*(1), 36–57. https://doi.org/10.1080/00087041.2017.1304498.

5. Gershon, N., & Page, W. (2001). What storytelling can do for information visualization. *Communications of the ACM, 44*(8), 31–37. https://doi.org/10.1145/381641.381653.

6. Hullman, J., Drucker, S., Riche, N. H., Lee, B., Fisher, D., & Adar, E. (2013). A deeper understanding of sequence in narrative visualization. *IEEE*

Transactions on Visualization and Computer Graphics, 19(12), 2406–2415. https://doi.org/10.1109/TVCG.2013.119.

7. Roth, R. E. (2020). Cartographic design as visual storytelling: Synthesis and review of map-based narratives, genres, and tropes. *The Cartographic Journal, 58*(1), 83–114. https://doi.org/10.1080/00087041.2019.1633103.

8. Singh, A., & Sharma, D. K. (2020). Image collection summarization: Past, present and future. In *Data Visualization and Knowledge Engineering* (pp. 49–78). Springer.

9. Chen, J., & Zhuge, H. (2019). Extractive summarization of documents with images based on multi-modal RNN. *Future Generation Computer Systems, 99*, 186–196.

10. Kuzovkin, D., Pouli, T., Cozot, R., Le Meur, O., Kervec, J., & Bouatouch, K. (2017). Context-aware clustering and assessment of photo collections. In *Proceedings of the Symposium on Computational Aesthetics* (pp. 1–10), Los Angeles, CA, USA, 29–30 July 2017.

11. Lotfi, F., Beheshti, A., Farhood, H., Pooshideh, M., Jamzad, M., & Beigy, H. (2023). Storytelling with image data: A systematic review and comparative analysis of methods and tools. *Algorithms, 16*(3), 135. https://doi.org/10.3390/a16030135.

12. Cao, R., Dey, S., Cunningham, A., Walsh, J., Smith, R. T., Zucco, J. E., & Thomas, B. H. (2020). Examining the use of narrative constructs in data videos. *Visual Informatics, 4*(1), 8–22. https://doi.org/10.1016/j.visinf.2019.12.002.

13. Aggarwal C. C., & Reddy C. K. (2014). *Data clustering: Algorithms and applications*. Chapman & Hall/CRC Data Mining and Knowledge Discovery Series.

14. Ahmed M. (2019). Data summarization: A survey. *Knowledge and Information Systems, 58*(1), 249–273. https://doi.org/10.1007/s10115-018-1243-0.

15. Mao J., Xu W., Yang Y., Wang J., Huang Z., & Yuille A. (2014). Deep captioning with multimodal recurrent neural networks (m-RNN). *arXiv preprint arXiv:1412.6632*. URL: https://arxiv.org/abs/1412.6632 (accesses May 16, 2025).

16. Karpathy A., & Fei-Fei L. (2015). Deep visual-semantic alignments for generating image descriptions. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3128–3137). https://doi.org/10.1109/CVPR.2015.7298932.

17. Vinyals O., Toshev A., Bengio S., & Erhan D. (2015). Show and tell: A neural image caption generator. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 3156–3164). https://doi.org/10.1109/CVPR.2015.7298935.

18. Rennie S. J., Marcheret E., Mroueh Y., Ross J., & Goel V. (2017). Self-critical sequence training for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 7008–7024). https://doi.org/10.1109/CVPR.2017.742.

19. Ren Z., Wang X., Zhang N., Lv X., & Li L.-J. (2017). Deep reinforcement learning-based image captioning with embedding reward. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 290–298). https://doi.org/10.1109/CVPR.2017.42.

20. Gordon D., Kembhavi A., Rastegari M., Redmon J., Fox D., & Farhadi A. (2018). IQA: Visual question answering in interactive environments. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 4089–4098). https://doi.org/10.1109/CVPR.2018.00431.

21. Patro B., Patel S., & Namboodiri V. (2020). Robust explanations for visual question answering. *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision* (pp. 1577–1586). https://doi.org/10.1109/WACV45572.2020.9093373.

22. Wu Q., Wang P., Shen C., Reid I., & Van Den Hengel A. (2018). Are you talking to me? Reasoned visual dialog generation through adversarial learning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6106–6115). https://doi.org/10.1109/CVPR.2018.00639.

23. Chen C., Mu S., Xiao W., Ye Z., Wu L., & Ju Q. (2019). Improving image captioning with conditional generative adversarial nets. *Proceedings of the AAAI Conference on Artificial Intelligence, 33*(01), 8142–8150. https://doi.org/10.1609/aaai.v33i01.33018142.

24. Xu K., Ba J., Kiros R., Cho K., Courville A., Salakhudinov R., Zemel R., & Bengio Y. (2015). Show, attend and tell: Neural image caption generation with visual attention. *Proceedings of the International Conference on Machine Learning* (pp. 2048–2057).

25. Bahdanau D., Cho K., & Bengio Y. (2014). Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*. URL: https://arxiv.org/abs/1409.0473 (accesses May 15, 2025).

26. Ba J., Mnih V., & Kavukcuoglu K. (2014). Multiple object recognition with visual attention. *arXiv preprint arXiv:1412.7755*. URL: https://arxiv.org/abs/1412.7755 (accesses May 19, 2025).

27. Mnih V., Heess N., & Graves A. (2014). Recurrent models of visual attention. *The 27th International Conference on Neural Information Processing Systems, 27*, 2204–2212.

28. Lu J., Xiong C., Parikh D., & Socher R. (2017). Knowing when to look: Adaptive attention via a visual sentinel for image captioning. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 375–383). https://doi.org/10.1109/CVPR.2017.47.

29. Anderson P., He X., Buehler C., Teney D., Johnson M., Gould S., & Zhang L. (2018). Bottom-up and top-down attention for image captioning and visual

question answering. *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition* (pp. 6077–6086). https://doi.org/10.1109/CVPR.2018.00636.

30. Qin Y., Du J., Zhang Y., & Lu H. (2019). Look back and predict forward in image captioning. *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition* (pp. 8367–8375). https://doi.org/10.1109/CVPR.2019.00857.

31. Uehara, K., Mori, Y., Mukuta, Y., & Harada, T. (2022). ViNTER: Image narrative generation with emotion-arc-aware transformer. *In Companion Proceedings of the Web Conference 2022* (pp. 716–725), Virtual Event/Lyon, France.