Comparison and Analysis of Digital Image Semantic Annotation Models: Postprint

Authors: Chen Jinju, Ou Shiyan

Date: 2023-08-26T00:00:00+00:00

Abstract

[Objective/Significance] The foundation of image semantic annotation lies in the construction of image semantic annotation models. Conducting a systematic review and summary of current mainstream image semantic annotation models, and analyzing their respective advantages and disadvantages in image semantic annotation, can provide valuable reference and guidance for future related research. [Method/Process] Through literature review, four major categories of image semantic annotation models are identified: the Eakins model, Jaimes&Chang model, Kong model, and Panofsky model. Subsequently, employing comparative and inductive methods, the first three models are analyzed and compared across three dimensions: semantic hierarchy, scalability, and application scope and modality. [Results/Conclusion] The Eakins model exhibits the most comprehensive semantic hierarchy, the strongest semantic expression capability, and the broadest application scope; the Kong model demonstrates the highest scalability and optimal adaptability.

Full Text

Preamble

Vol. 62 No. 6 March 2018 ChinaXiv Cooperative Journal

Comparison and Analysis of Digital Image Semantic Annotation Models

Chen Jinju, Ou Shiyan
School of Information Management, Nanjing University, Nanjing 210023

Abstract

[Purpose/Significance] The foundation of image semantic annotation lies in the construction of semantic annotation models. By systematically reviewing

and summarizing current mainstream image semantic annotation models and analyzing their respective advantages and disadvantages in image semantic annotation, this study provides valuable references for future research. [**Method/Process**] Through literature review, we identified four major categories of image semantic annotation models: the Eakins model, the Jaimes & Chang model, the Kong model, and the Panofsky model. We then employed comparative and inductive methods to analyze the first three models from three perspectives: semantic levels, extensibility, and application scope and methods. [**Result/Conclusion**] The Eakins model demonstrates the most comprehensive semantic hierarchy, strongest semantic expression capability, and widest application range, while the Kong model exhibits the greatest extensibility and adaptability.

Keywords: image annotation; semantic image annotation; semantic models for image annotation
Classification Number: G250
DOI: 10.13266/j.issn.0252-3116.2018.06.014

Introduction

With the rapid development of digital imaging and multimedia technologies in recent years, massive amounts of digital image resources have emerged, making effective retrieval and utilization of these resources an increasingly urgent challenge. Early image annotation primarily relied on manual selection of subject terms or keywords to describe image content. Although this approach achieved relatively high accuracy, it was labor-intensive and often yielded subjective and unstable results. As computer technology advanced, content-based automatic image annotation gradually became mainstream. Such annotation approaches image retrieval by automatically extracting low-level visual features (e.g., color, shape, texture) from images and matching them with user semantic queries. However, since visual features alone cannot fully reflect user search intent, this has led to the well-known "semantic gap" problem [1]. To facilitate sharing and reuse of digital image resources, effective semantic annotation of image content is essential to enhance human understanding of images. Image semantic annotation heavily depends on image semantic annotation models, which are conceptual models abstracted from image content. These models typically employ hierarchical structures to describe visual features (color, texture, shape), logical features (e.g., contained objects and their relative relationships), and semantic features (e.g., scenes, emotions) at different levels from low to high, thereby helping people better understand and describe image content. Due to this hierarchical structure, such models are often referred to as image hierarchy models.

Researchers from various fields—including computer science, biomedicine, and library and information science—have studied image semantic annotation, particularly image semantic annotation models. To gain a comprehensive understanding of the research landscape, we retrieved over 40 Chinese and English

publications on image semantic annotation models from 1985 to 2017 in the Web of Science, CNKI, and Google Scholar databases. Analysis revealed that while these publications mention more than 20 different models, most derive from four fundamental models: the Eakins model [2], the Jaimes & Chang model [3], the Kong model [26], and the Panofsky model [29], belonging to four distinct families as shown in Table 1 . In addition to these mainstream models, there are some less widely applied models, such as M.G. Krause's two-layer image content framework [7], B. Burford et al.'s six-layer model [8], and Y. Badr and R. Chbeir's two-layer model [9]. Due to their limited application scope, these are not considered mainstream image semantic annotation models. Although research on image semantic annotation models is abundant, comprehensive and systematic reviews and analyses of these models remain scarce.

This paper begins with the four fundamental image semantic annotation models, providing a comprehensive review and comparative analysis of mainstream models to offer references for the construction and application of image semantic annotation models.

2. The Eakins Image Semantic Hierarchy Model and Its Derivatives

2.1 The Retrieval-Oriented Eakins Image Semantic Hierarchy Model

In 1996, British scholar J.P. Eakins proposed a simple yet practical hierarchical image semantic model (hereinafter referred to as the "Eakins model") [10-12]. From the perspective of retrieval needs, Eakins first divided image semantic content into three basic levels from bottom to top: the primitive features level, the object level, and the semantic concept level, with each level further subdivided into finer-grained categories as shown in Figure 1 Figure 1: see original paper.

Primitive Features Level: Primitive features refer to objective, purely visual features directly derived from the image itself, which can be categorized into five types: color features, texture features, shape features, spatial location features, and combinations of these four features. None involve image semantic content. Image retrieval at this level requires no external knowledge reference. Content-based image retrieval primarily operates at this level and is widely used in various specialized image retrieval applications, such as trademark image retrieval during trademark registration processes [13].

Logical Features Level: Logical features are derived through logical reasoning about objects in an image based on its visual features, involving image semantic content. Image retrieval based on logical features can be further subdivided into: retrieving images of specific object types and retrieving images of individual objects or persons. Retrieval at this level is more universally applicable than semantic retrieval at the primitive features level.

Abstract Attributes Level: Abstract attributes are features obtained through abstraction and subjective reasoning about the purpose and meaning

of object scenes in an image. Image retrieval based on abstract attributes can be further subdivided into: retrieving images of specific events or activities and retrieving images with emotional or symbolic meanings. Retrieval at this level requires not only understanding of image semantic content and background knowledge but also certain reasoning and judgment capabilities.

The Eakins model was among the earliest image semantic models, designed primarily for image retrieval rather than image sharing and reuse. Before its proposal, scholars V.N. Gudivada and V.V. Raghavan from East Carolina University and the University of Louisiana at Lafayette had proposed a two-layer image semantic model comprising primitive and logical features levels in 1995 [14]. This model served as the foundation for the Eakins model but had weaker semantic expression capability due to its lack of abstraction and description of object scenes.

2.2 Fine-Tuning and Reinterpretation of the Eakins Model for Semantic Annotation

Following the proposal of the Eakins model, domestic scholars (e.g., Wu Renjie [15], Zhang Jie [16], Lu Quan et al. [17]) reinterpreted, elaborated on, and fine-tuned it to make it applicable to image semantic annotation. These improvements essentially re-interpreted the Eakins model as the three-level structure shown in Figure 1 Figure 1: see original paper [15-18]:

Low-Level Features Level: This level corresponds to the primitive features level of the Eakins model, reflecting no image semantic content information. However, it removes spatial location features from the primitive features level, retaining only color, texture, and shape features.

Object Level: This level combines the logical features level and part of the primitive features level of the Eakins model, subdivided into two sub-levels: object semantics level and object space level. The former contains specific objects involved in the image, corresponding to the logical features level of the Eakins model; the latter refers to the spatial relationships between objects, corresponding to the spatial location features in the Eakins model's primitive features level.

Semantic Concept Level: This level corresponds to the abstract attributes level of the Eakins model but emphasizes scene semantics description. It is subdivided into three sub-levels: scene semantics level (the environment where objects are located), behavior semantics level (the activities of objects), and emotional semantics level (subjective feelings evoked by the image).

The Eakins model is an integrated whole with dependencies between levels, where middle and high-level semantics are typically obtained based on low-level features through prior knowledge and reasoning. However, efficient extraction of high-level semantics remains challenging. The reinterpreted Eakins model does not differ essentially from the original but features clearer, more explicit

semantic levels with greater universality, which explains its widespread adoption in domestic image semantic annotation research. All subsequent references to the Eakins model in this paper denote this fine-tuned and reinterpreted version.

2.3 Derivative Models of the Eakins Model

After the creation of the Eakins model, researchers developed numerous derivative versions by adding, removing, or adjusting semantic levels for specific applications. In 1998, D. Hong et al. made local adjustments to the Eakins model's sub-levels, dividing image content into three levels: basic visual content, object content, and scene content [19], corresponding to the low-level features, object semantics, and scene semantics levels of the Eakins model. This adjustment aimed to flexibly describe images for specific retrieval scenarios but suffered from treating scenes as global descriptions without considering behavior and emotional semantics, resulting in weaker semantic expression capability than the Eakins model.

Domestic scholar Yu Yongxin argued that the semantic gap primarily resulted from insufficient description of relationships between entities in images. Therefore, he elevated two sub-levels from the Eakins model's object level—object semantics level and object space level—to become primary levels alongside low-level features and semantic concept levels, deriving a four-layer model [20]. Although this model expressed no essentially different semantics from the Eakins model, it emphasized description of entity relationships.

Additionally, some researchers expanded the Eakins model's semantic hierarchy to create multi-layer models with richer semantics. Cai Changxu and Peng Yang proposed their respective seven-layer semantic models in 2005 and 2007. These models share essentially the same levels: the first six correspond to the six sub-levels of the Eakins model, while the seventh level expresses higher-level semantics of the image—abstract and real-world understanding, typically referring to the real scenario reflected in the image (e.g., a wedding, the Hongmen Banquet) [4,21]. Compared with the first six levels, this higher-level semantics layer focuses on describing the global connotation of the image from an overall perspective with greater abstraction. Although these two models have no essential difference, they apply to different image types: the former targets general static images, while the latter focuses on animation material images.

3. The Jaimes & Chang Image Semantic Hierarchy Model and Its Derivatives

3.1 The Jaimes & Chang Image Semantic Hierarchy Model

In 1998, A. Jaimes and S.F. Chang proposed a visual information classification framework for automatic image classification, which 自下而上 contained five levels: region, perception, object part, object, and scene [22]. In 2000, the authors integrated knowledge from multiple domains (such as art and cognitive

psychology) to transform this framework into a conceptual framework for image indexing (hereinafter referred to as the "Jaimes & Chang model") [3]. In this framework, image content is divided into non-visual and visual categories, as shown in Figure 2 [Figure 2: see original paper].

Non-visual content refers to information closely related to but not directly part of the image, including physical attributes, catalog information, and related information. **Visual content** refers to information directly perceived when observing an image, which 自下而上 forms a pyramid structure of ten levels: type technology, global distribution, local structure, global composition, generic objects, generic scene, specific objects, specific scene, abstract objects, and abstract scene [3,23]. The first four levels describe image syntax or perception, involving features such as color, texture, and spatial layout perceived by humans or machines, which are the primary focus of content-based image annotation and retrieval. The remaining six levels describe semantics or visual concepts, which are the main concern of semantic image annotation and retrieval. Compared with the 1998 visual information classification framework, this model pays greater attention to object semantics and scene semantics, providing more fine-grained descriptions of objects and scenes from generic, specific, and abstract perspectives. However, it is less suitable for images where objects and scenes are not the primary annotation focus.

3.2 Improvements and Adjustments to the Jaimes & Chang Model

The Jaimes & Chang model is a relatively universal semantic annotation model that scholars have improved to create derivative models for specific applications. In 2004, L. Hollink et al. addressed the mismatch between user needs and existing image retrieval technologies by locally adding, removing, and adjusting sub-levels of the Jaimes & Chang model, making the model's description granularity coarser to encompass more semantic content. They created a user image description classification framework 自下而上 comprising three levels: conceptual, perceptual, and non-visual. The conceptual level corresponds to the semantic/visual concept level of the Jaimes & Chang model but adds time, location, and events. The perceptual level roughly corresponds to the syntax/perception level of the Jaimes & Chang model but without further sub-levels. The non-visual level roughly corresponds to the non-visual content level of the Jaimes & Chang model, focusing on descriptive image metadata such as creator, date, and title [24]. This framework does not extend the semantic hierarchy of the Jaimes & Chang model but merely adjusts the granularity and content of its sub-levels.

In 2011, E.K. Chung and J.W. Yoon proposed an image feature description framework based on in-depth analysis of the underlying structure of image retrieval needs, 自上而下 containing three levels: non-visual features, syntactic features, and semantic features [5]. Non-visual features roughly correspond to the non-visual content level of the Jaimes & Chang model but expand related information to contextual information. Syntactic features roughly correspond to

the syntax/perception level of the Jaimes & Chang model but remove type technology. Semantic features roughly correspond to the semantic/visual concept level of the Jaimes & Chang model but add people, time, location, and activities. This model similarly only adjusts the granularity and content within the Jaimes & Chang model but extends its semantic/visual concept level, resulting in enhanced overall semantic expression capability.

In the same year, the same authors found that questions expressed in natural language could better reflect users' image retrieval needs. Therefore, they adjusted the three-level image feature description framework into three levels: image need features, image features, and related information [25]. Image need features refer to the contextual environment of user image retrieval needs, such as retrieval motivation. The image features level contains non-visual features, syntax/perception object features, and semantic/visual concept features from the Jaimes & Chang model. The related information level separates the related information from the non-visual features of the Jaimes & Chang model as an independent level. This model adds an image need features level based on the Jaimes & Chang model's semantic hierarchy, enhancing overall semantic expression capability but also increasing model complexity and annotation difficulty.

In 2006, J.S. Hare et al. attempted to bridge the semantic gap in image retrieval by characterizing it as the transition from raw media (images) to full semantic understanding of media content (object relationships and beyond). They proposed a five-level semantic gradient model comprising raw image, visual descriptor, object, object name, and semantics [6]. Compared with the Jaimes & Chang model, this model has coarser semantic level granularity, lacks fine-grained descriptions of objects and scenes, and is suitable for image annotation with relatively coarse semantic granularity.

4. The Kong Image Semantic Annotation Model and Related Models

4.1 The Kong Image Semantic Annotation Model

In 2006, H. Kong et al. approached image content classification from the perspective of objects contained in images and proposed an extensible image semantic annotation ontology model (hereinafter referred to as the "Kong model") [26]. This model first includes a top-level ontology defining seven classes for describing object types in images: person, animal, plant, artifact, food, natural object, and natural phenomenon. This provides a generic image annotation framework that comprehensively covers semantic content at the object level, though with relatively coarse semantic granularity. To represent spatial relationships between objects and between objects and background, the model also defines a spatial ontology containing eight directional relations and eight topological relations.

Furthermore, to achieve higher precision in image retrieval, H. Kong et al. used personalized ontologies for image semantic annotation, allowing users to build

personalized ontologies based on the top-level ontology according to their needs. Figure 3 [Figure 3: see original paper] illustrates the transformation process from top-level ontology to personalized ontology, demonstrating how external knowledge about specific objects can be added to the top-level ontology to create a personalized ontology. For example, to semantically annotate an image of basketball player Yi Jianlian, a user might possess external knowledge such as "Yi Jianlian is a Chinese male athlete who plays for the Guangdong Hongyuan Basketball Club." The user would first locate the corresponding classes "natural object" and "artifact" in the top-level ontology for the object's nationality and club, then establish associations between these class instances ("China" and "Guangdong Hongyuan") and the object "Yi Jianlian," generating a personalized ontology for Yi Jianlian.

4.2 Related Models of the Kong Model

Some scholars have borrowed H. Kong et al.'s object-centered image content classification approach and the idea of using personalized ontologies for semantic annotation to construct ontology-based image semantic annotation and retrieval models. In 2008, Deng Tao et al. proposed an ontology-based image semantic annotation and retrieval model called ImageQ, which includes a four-layer image content description model. 自上而下, this model comprises: image metadata (reflecting external image features); object and its background and scene information; subject object and its related attributes; and semantic relationships between subject and object [27]. Compared with the Kong model, this model still focuses on object-level semantics but expands semantics related to objects and external image features, resulting in stronger semantic expression capability. Following Kong et al.'s approach, Deng Tao et al. provided only a top-level generic ontology model for specific application domains but allowed users to update the top-level ontology by adding, modifying, and deleting concepts, properties, and relationships to ultimately achieve domain ontology personalization.

In 2010, Shi Tingting et al. borrowed Deng Tao et al.'s four-layer image content description model to propose a three-layer image content description model [28]. Compared with Deng Tao et al.'s four-layer model, this model removed the image metadata features level, focusing only on internal semantic features rather than external image features. However, it adopted a similar object-centered approach to building personalized ontologies, which effectively enhanced the flexibility of image retrieval systems.

5. The Panofsky Image Semantic Hierarchy Model and Its Extensions

5.1 The Panofsky Image Semantic Hierarchy Model

In 1955, E. Panofsky proposed an analytical model (hereinafter referred to as the "Panofsky model") while studying Renaissance art images, as shown in

Figure 4 [Figure 4: see original paper]. The model comprises three levels: pre-iconographic description (referring to descriptions of themes expressed in images, including facts and emotions), iconographic analysis (referring to analysis of objectively identifiable items in images), and iconological interpretation (referring to interpretation of image connotation) [29]. This model primarily focuses on high-level semantic information in images without considering low-level physical features. Notably, this model is a theoretical analytical framework rather than a concrete semantic annotation model and cannot be directly applied to image semantic annotation.

5.2 Extensions of the Panofsky Model

To apply the Panofsky model to concrete image annotation, some scholars have extended it to create models with richer semantics. In 1986, S. Shatford horizontally expanded the three semantic levels of the Panofsky model, proposing a two-dimensional model known as the Panofsky-Shatford facet matrix, intended for application to all types of image semantic annotation [30-31]. This model contains three levels—generic, specific, and abstract—corresponding to the three levels of the Panofsky model (pre-iconographic, iconographic, and iconological). Each level includes four sub-levels: who, what, when, and where, corresponding to objects, events (activities), time, and location in images, forming 12 categories of image features that greatly enrich and refine the semantic content of the Panofsky model [32]. For example, abstract location represents symbolic places (e.g., heaven). This matrix was later applied to image indexing by N. Conduit and P. Rafferty, who further refined it based on user queries in image databases and 33 image features that archivists focused on in their work, such as subdividing generic location into indoor and outdoor [33]. These refinements made the Panofsky-Shatford facet matrix more comprehensive and complete.

In 2007, P. Rafferty and R. Hidderley borrowed methods from the Panofsky model and others to interpret images and proposed a six-layer model from the perspective of image content indexing, comprising bibliographic information, structural content, whole content, object content, interpretation of the whole image, and interpretation of objects within it [34]. This model expands external physical features of images, focusing on middle and high-level semantic features with emphasis on object and emotion description.

In 2013, F. Fauzi and M. Belkhatir borrowed from the Panofsky model and other related models to propose a user-centered, concept-based framework for automatic multifaceted indexing. This framework analyzed the semantics of Web image contextual information and divided it into five broad semantic concepts: (1) Signals: referring to low-level visual features; (2) Objects: referring to entities in images, divided into animate and inanimate; (3) Relations: referring to relationships between objects in images and external feature relationships such as creator and image type; (4) Scenes: referring to describing the image as a whole based on all contained objects; and (5) Abstractions: referring to abstract concepts expressed in images [35]. This framework extends the Panof-

sky model, which originally contained only middle and high-level semantics, by adding descriptions of low-level visual features (signals) and vertically expanding the iconographic level with object and relation levels, thereby enhancing the framework's overall semantic expression capability.

6. Comparative Analysis of Models

Based on the preceding discussion of the four image semantic annotation models, this section conducts an in-depth comparison of the Eakins model, Jaimes & Chang model, and Kong model. Since the Panofsky model is merely a theoretical analytical framework rather than a concrete model, it is excluded from this comparative analysis. Given the numerous derivative models within each family, we selected only the original basic model from each category as the object of analysis, examining them from three aspects: semantic levels, extensibility, and application scope and methods. Two important metrics for evaluating image semantic annotation models are: (1) semantic expression capability—whether and to what extent the model can completely express semantics contained in images; and (2) adaptability—whether the model can meet different user needs, with an important method for improving adaptability being allowing users to extend the model during application.

6.1 Semantic Levels

The more comprehensive the semantic levels of an image semantic annotation model, the richer the semantics it can express. Image content features can be divided into three major categories: physical features, object features, and semantic features. Physical features do not involve image semantic content, while object features and semantic features contain five types of image semantic content: object semantics, object space, scene semantics, behavior semantics, and emotional semantics. Both the Eakins model and the Jaimes & Chang model include image content features from three levels—low-level physical features, object features, and high-level semantic features—expressing increasingly abstract semantics from low to high. However, they differ in the number of semantic features they contain. The Eakins model includes five major semantic features (object semantics, object space, scene semantics, behavior semantics, and emotional semantics), making its semantic hierarchy the most comprehensive with the most complete semantic expression. The Jaimes & Chang model includes only three major semantic features (object semantics, scene semantics, and emotional semantics), lacking descriptions of object space and behavior semantics, making its semantic comprehensiveness and completeness weaker than the Eakins model. The Kong model does not consider low-level physical features or high-level semantic features, focusing directly on middle-level object features to describe objects and their spatial relationships in images. This model contains two object features—object semantics and object space—making its semantic hierarchy comprehensiveness and expression completeness weaker than both the Eakins and Jaimes & Chang models, as shown in Table 2 .

Table 2. Semantic Levels of Image Semantic Annotation Models

Model	Object Semantics	Object Space	Scene Semantics	Behavior Semantics	Emotional Semantics
Eakins Model	P	P	P	P	P
Jaimes & Chang Model	P	-	P	-	P
Kong Model	P	P	-	-	-

Note: "P" indicates that the model includes the corresponding semantic level.

In summary, the Eakins model has the most complete semantics and strongest expression capability. Human understanding of image semantics, derived from comprehension and cognition, often focuses on object semantics and high-level semantics. While the core models considered in this paper all incorporate this factor, they lack description of more abstract semantics such as atmosphere rendered by images, leaving room for further extension beyond the semantic concept level.

6.2 Extensibility

The extensibility of image semantic annotation models significantly impacts the vitality and adaptability of image retrieval systems. We summarize the extensibility of the three core models as strong, relatively strong, and weak, as shown in Table 3 . All three core models allow user extension. The Eakins model supports addition/removal of semantic levels and fine-tuning of semantic hierarchies, while the Kong model supports adding sub-levels at the object semantics level. Both thus exhibit good adaptability. Although the Jaimes & Chang model also allows user extension, it primarily supports adding semantic levels and hierarchies but not incorporating external knowledge during application, resulting in relatively weaker extensibility and adaptability.

Table 3. Extensibility of Image Semantic Annotation Models

Model	Extensibility	Application Reference to External Knowledge
Eakins Model	Strong	Yes
Jaimes & Chang Model	Relatively Strong	No
Kong Model	Strong	Yes

Furthermore, both the Eakins model and Kong model allow incorporation of external knowledge (such as contextual information) during image semantic anno-

tation to enrich image semantics, rather than being limited to content reflected in the image itself. The Kong model also employs ontology technology, allowing users to build personalized ontologies based on the top-level ontology according to domain knowledge, further enhancing model extensibility and adaptability.

Applying ontology technology to image semantic annotation offers several advantages: ontology is a standardized, normalized knowledge representation method that provides unified concepts and relationships within a domain [36]; ontology establishes detailed descriptions of concepts and relationships, creating connections between dispersed and isolated images and enhancing image coupling [27]; ontology provides semantic description methods independent of specific objects, facilitating sharing and reuse of semantic information [37-38]; and ontology's reasoning function enables intelligent image retrieval. Different ontologies can be used to describe images for different applications. For example, low-level image features can be described using the VDO (visual descriptor ontology) ontology, which includes MPEG-7 visual descriptors and concepts and attributes of object visual features [39]; middle and high-level semantic information typically requires domain-specific ontologies. Given the Kong model's extensibility and support for personalized customization based on external knowledge, it demonstrates the strongest extensibility and adaptability.

6.3 Application Scope and Methods

All three core models can be used for image semantic annotation and retrieval, with no specific restrictions on applicable image types or application domains. By analyzing the semantic information described by the models, we can summarize the characteristics of applicable images and application scenarios for each core model, as shown in Table 4 . Although the three models are all image semantic annotation models, their application scopes differ.

Table 4. Application Scope of Image Semantic Annotation Models

Model	Applicable Image Characteristics	Application Scenarios
Eakins Model	Comprehensive and diverse semantic content, including objects, object space, behavior, scene, and emotion	Image semantic annotation in art (painting, calligraphy, etc.), history, and other fields
Jaimes & Chang Model	Primarily object and scene semantics, generally excluding emotional semantics	Image annotation in architecture design, geography, and other fields
Kong Model	Primarily object and object space semantics, generally excluding emotional semantics	Image annotation in biomedicine, healthcare, and other fields

The Eakins model is primarily used for image retrieval, aiming to improve retrieval system performance and precision. This model can comprehensively express objects, object space, behavior, scene, and emotional semantics in images, applicable to but not limited to image semantic annotation in art (painting, calligraphy, etc.) and history fields. For example, Wang Xiaoguang and Xu Lei proposed a semantic description hierarchy model for Dunhuang mural digital images in 2014 [40]. Based on the Eakins model, this model incorporated terminology and image metadata suitable for Dunhuang mural description to reveal high-level image semantics, achieving semantic annotation of Dunhuang mural digital images. In 2017, Xu Lei and Wang Xiaoguang combined narrative image plot semantics with their digital image semantic description hierarchy model to model plot semantics in narrative images, achieving semantic annotation and retrieval of such images [41]. Narrative images contain relatively comprehensive semantic information that basically encompasses object, object space, behavior, scene, and emotional semantics.

The Jaimes & Chang model is primarily used for retrieval-oriented indexing and classification of image descriptions. In 2001, C. Jørgensen et al. conducted an exploratory evaluation of the Jaimes & Chang pyramid model, validating it through application to image semantic description and indexing. Results showed the model to be powerful, capable of describing visual content for retrieval, guiding the indexing process, and classifying manually or automatically obtained descriptions, effectively covering image features involved in user description and annotation processes [42]. The Jaimes & Chang model emphasizes fine-grained description of object semantics and scene semantics, making it suitable for images focusing on objects and scenes, with applications in architecture design, geography, and other fields. However, since this model does not include emotional semantics, it is not suitable for describing emotion-laden images.

The Kong model is also primarily used for image retrieval, typically building models from the image object semantics level. It employs ontology technology to construct a top-level ontology and allows users to build personalized ontologies according to their needs. Deng Tao and Shi Tingting et al. conducted experiments in image retrieval systems, demonstrating that image semantic annotation and retrieval based on personalized ontologies achieved higher precision compared to other retrieval methods (such as Baidu image search using search engines) [33-34]. The Kong model is an object-level semantic annotation model that does not consider physical or semantic features. It has strong descriptive capability for images with objective objects as main content but generally does not include emotional semantics, making it applicable to but not limited to biomedicine, healthcare, and other fields.

Conclusion

Image semantic annotation models provide a prerequisite and foundation for image semantic annotation and retrieval, offering a framework for describing image content (including low-level visual features and semantic features). This

paper analyzed and summarized four major image semantic annotation models (the Eakins model, Jaimes & Chang model, Kong model, and Panofsky model) and their derivatives. Among these four models, except for the Panofsky model, which is an abstract image semantic analysis framework, the remaining three are practically applicable models. Therefore, we compared and analyzed these three models from three aspects: semantic levels, extensibility, and application scope and methods. In terms of semantic expression capability, the Eakins model has the most comprehensive semantic hierarchy and strongest semantic expression capability. In terms of adaptability, the Kong model demonstrates the best adaptability, not only referencing external knowledge during annotation but also allowing users to extend the model during application and build personalized ontologies according to their expertise. In terms of application scope, the Eakins model is the most widely applied, with many researchers proposing numerous improved models and related applications based on it, leading to broad recognition. Although the other two models have also been widely applied, their influence is relatively weaker compared to the Eakins model.

References

[1] Smeulders AWM, Worring M, Santini S, et al. Content-based image retrieval at the end of the early years[J]. IEEE transactions on pattern analysis and machine intelligence, 2000, 22(12): 1349-1379.

[2] Eakins JP. Retrieval of still images by content[M]. Lectures on information retrieval. Springer, Berlin, Heidelberg, 2000: 111-130.

[3] Jaimes A, Chang SF. A conceptual framework for indexing visual information at multiple levels[C]//Proceeding of SPIE-The International Society for Optical Engineering. San Jose: IS&T/SPIE Internet imaging, 2000, 3964: 2-16.

[4] Cai Changxu. Research on semantic-based image annotation and retrieval system[D]. Wuhan: Wuhan University, 2005.

[5] Chung EK, Yoon JW. Image needs in the context of image use: an exploratory study[J]. Journal of information science, 2011, 37(2): 163-177.

[6] Hare JS, Lewis PH, Enser PGB, et al. Mind the gap: another look at the problem of the semantic gap in image retrieval[J]. Multimedia Content Analysis Management & Retrieval, 2006, spie v.

[7] Krause MG. Intellectual problems of indexing picture collections[J]. Audiovisual librarian, 1988, 14(2): 73-81.

[8] Burford B, Briggs P, Eakins JP. A taxonomy of the image: on the classification of content for image retrieval[J]. Visual communication, 2003, 2(2): 123-161.

[9] Badr Y, Chbeir R. Automatic image description based on textual data[M]//Journal on data semantics VII. Berlin, Heidelberg: Springer, 2006.

[10] Eakins JP. Automatic image content retrieval—are we getting anywhere?[C]//Proceeding of Third International Conference on Electronic Library and Visual Information Research. De Montfort University. Milton Keynes: Aslib, 1996: 123-125.

[11] Eakins JP. Design criteria for a shape retrieval system[J]. Computers in industry, 1993, 21(2): 167-184.

[12] Eakins JP, Graham ME, Boardman JM, et al. Retrieval of trademark images by shape feature[C]//Proceeding of first International conference on electronic library and visual information system research. Milton Keynes: De Montfort University, 1996: 101-109.

[13] Petkovic D. Query by image content[C]//Oral presentation to storage and retrieval for image and video databases. California: San Jose, 1996.

[14] Gudivada VN, Raghavan VV. Content-based image retrieval systems[J]. IEEE computer, 1995, 28(9): 18-22.

[15] Wu Renjie. Preliminary study on hierarchical semantic description of images[J]. Computer development and application, 2011(5): 12-14.

[16] Zhang Jie. Image semantic annotation[J]. Computer development and application, 2012(1): 10-12.

[17] Lu Quan, Ding Heng. Review of emotion-based image retrieval research[J]. Information theory and practice, 2013(2): 119-124.

[18] Huang Zhichun. Research on semantic-based image retrieval and related technologies[D]. Guangzhou: South China University of Technology, 2012.

[19] Hong D, Wu J, Singh SS. Refining image retrieval based on context-driven methods[C]//Storage and retrieval for image and video databases VII. 1998: 581-592.

[20] Yu Yongxin. Research on ontology-based image semantic recognition and retrieval[D]. Tianjin: Tianjin University, 2009.

[21] Peng Yang. Research on ontology-based semantic annotation of animation material images[D]. Changsha: Hunan Normal University, 2009.

[22] Jaimes A, Chang SF. Model-based classification of visual information for content-based retrieval[C]//Proceedings of SPIE-The International Society for Optical Engineering. 1998: 402-413.

[23] Tousch AM, Herbin S, Audibert JY. Semantic hierarchies for image annotation: a survey[J]. Pattern recognition, 2012, 45(1): 333-345.

[24] Hollink L, Schreiber AT, Wielinga BJ, et al. Classification of user image descriptions[J]. International journal of human-computer studies, 2004, 61(5): 601-626.

[25] Yoon JW, Chung EK. Understanding image needs in daily life by analyzing questions in a social Q&A site[J]. Journal of the Association for Information Science & Technology, 2011, 62(11): 2201-2213.

[26] Kong H, Hwang M, Kim P. The study on semantic image retrieval based on the personalized ontology[J]. International journal of information technology, 2006, 12(2): 35-46.

[27] Deng Tao, Guo Lei, Yang Weili. Ontology-based image semantic annotation and retrieval model[J]. Computer engineering, 2008(17): 188-190.

[28] Shi Tingting, Yan Dashun, Shen Yuli. Image semantic annotation and retrieval based on personalized ontology[J]. Computer application, 2010(1): 90-93.

[29] Panofsky E. Meaning in the visual art: papers in and on art history[M]. New York: Doubleday Anchor Books, 1955: 39-40.

[30] Shatford S. Analyzing the subject of a picture: a theoretical approach[J]. Cataloging & classification quarterly, 1986, 6(3): 39-62.

[31] Huang?, Wang Shanshan, Geng Qian. Progress and implications of foreign image feature research[J]. Library and information service, 2015, 59(8): 138-146.

[32] Choi Y, Rasmussen EM. Searching for images: the analysis of users' queries for image retrieval in American history[J]. Journal of the Association for Information Science and Technology, 2003, 54(6): 498-511.

[33] Conduit N, Rafferty P. Constructing an image indexing template for the children's society: users' queries and archivists' practice[J]. Journal of documentation, 2007, 63(6): 898-919.

[34] Rafferty P, Hidderley R. Flickr and democratic indexing: dialogic approaches to indexing[J]. Aslib Proceedings, 2007, 59(4/5): 397-410.

[35] Fauzi F, Belkhatir M. Multifaceted conceptual image indexing on the World Wide Web[J]. Information processing & management, 2013, 49(2): 420-440.

[36] Zhang Mei, Hao Jia, Yan Yan, et al. Ontology-based knowledge modeling technology[J]. Journal of Beijing Institute of Technology, 2010(12): 1405-1408, 1431.

[37] Zhang Yang, Fang Bin, Xu Chuanyun. Image semantic recognition based on ontology and description logic[C]//Nanning: National Conference on Safety Critical Technology and Application. 2009.

[38] Brachman RJ, Schmolze JG. An overview of the KL-ONE knowledge representation system[J]. Cognitive science, 1985, 9(2): 171-216.

[39] Simou N, Tzouvaras V, Avrithis Y, et al. A visual descriptor ontology for multimedia reasoning[C]//Proceedings of workshop on image analysis for multimedia interactive services. Montreux, 2005: 13-15.

[40] Wang Xiaoguang, Xu Lei, Li Gang. Research on semantic description methods for Dunhuang mural digital images[J]. Journal of Library Science in China, 2014, 40(1): 50-59.

[41] Xu Lei, Wang Xiaoguang. Research on narrative image semantic annotation model[J]. Journal of Library Science in China, 2017, 43(5): 70-83.

[42] Jørgensen C, Jaimes A, Benitez AB, et al. A conceptual framework and empirical research for classifying visual descriptors[J]. Journal of the Association for Information Science and Technology, 2001, 52(11): 938-947.

Author Contributions: Chen Jinju: wrote and revised the paper; Ou Shiyan: proposed the research direction and outlined key points, revised the paper.

Comparison and Analysis of the Semantic Models for Digital Image Annotation

Chen Jinju, Ou Shiyan
School of Information Management, Nanjing University, Nanjing 210023

Abstract: [Purpose/significance] Semantic annotation of digital images is an effective way to solve the problem of image retrieval. The foundation of semantic image annotation is the construction of semantic models. This paper intends to review the existing mainstream semantic models for image annotation, and explore their advantages and disadvantages. [Method/process] Firstly, four representative semantic models for image annotation were reviewed, including Eakins model, Jaimes & Chang model, Kong model and Panofsky model, using literature survey, and then the first three models from three aspects (i.e. semantic level, extensibility and application range) were compared and analyzed using comparative analysis. [Result/conclusion] Through the above analysis, it can be concluded that Eakins model has the most comprehensive semantic level, the strongest semantic expression ability and the widest application range, whereas Kong model is the most scalable and adaptable one.

Keywords: image annotation; semantic image annotation; semantic models for image annotation

Note: Figure translations are in progress. See original paper for figures.