ISSN 2007-9737

# Analysis of Automatic Annotations of Real Video Surveillance Images

Diana Karina Guevara Flores[1], Fernando Pérez Téllez[2], David Eduardo Pinto Avendaño[1]

[1] Benémerita Universidad Autónoma de Puebla,
Department of Computing,
México

[2] Technological University Dublin,
Department of Computing,
Ireland

n.edd.a@hotmail.com

**Abstract.** The results of the analysis of the automatic annotations of real video surveillance sequences are presented. The annotations of the frames of surveillance sequences of the parking lot of a university campus are generated. The purpose of the analysis is to evaluate the quality of the descriptions and analyze the correspondence between the semantic content of the images and the corresponding annotation. To perform the tests, a fixed camera was placed in the campus parking lot and video sequences of about 20 minutes were obtained, later each frame was annotated individually and a text repository with all the annotations was formed. It was observed that it is possible to take advantage of the properties of the video to evaluate the performance of the annotator and the example of the crossing of a pedestrian is presented as an example for its analysis.

**Keywords.** Automatic annotation, semantic analysis, surveillance images.

## 1 Introduction

The amount of images and video available on the Internet increases exponentially, however, it is mostly non-structured data, because of this, there is a large number of automatic annotators [13, 17, 5] that allow the generation of metadata of images, a fundamental step towards the analysis and processing of large amounts of non-structured data.

However these automatic annotators have not reached a level of robustness that achieves the performance of human beings. Additionally, a notable performance have been obtained with databases such as MS COCO [12] or Flickr [15], however, tests with real data have not achieved sufficiently accurate results, mainly when the test environment differs from the domain of the database used in the training.

Corporations like *Amazon*, *Google* and *Microsoft* are developing systems and services oriented to this task, and although they have achieved remarkable performance in tasks such as the detection of faces or objects, the annotation of free domain images, such as those from social networks, remains a challenging task.

Among the reasons why automatic annotator results are still far from achieving human performance, it is found that traditional machine learning techniques do not allow the processing of heterogeneous input data and deep learning techniques require large databases for their training, and the existing ones are still insufficient.

The annotation of free domain images allows the use of natural language processing tools in the analysis of images and video, with which a large number of problems, such as the recognition of objects and actions, could be addressed with techniques other than the traditional ones.

There are also open source automatic annotators of images with which it is possible to perform prototypes of systems oriented to the execution of tasks of higher level, such as the detection of anomalies.

Therefore, it is proposed to carry out a first approach to the analysis of automatically generated annotations of video surveillance sequences where each frame is processed individually and analyze the performance of a open source automatic annotator.

## 2 Problem Description

The goal of automatic image annotation is to assign a collection of words to an objective image that describes it in the best possible way. Automatic image annotation has aroused the interest of the scientific community due to the wide range of tasks with which it is related, such as image recovery, automatic elaboration of abstracts or indexing.

Automatic annotation is a difficult task because the same object can be captured from different angles, distances or lighting conditions. In addition, objects with the same name can have many variations of color, shape or texture.

Automatic image annotation is a key step towards the image recovery based on semantic keywords, which is a widely used way to recover images on the web [9, 7, 20]. The increase in the number of images in social networks increases the demand for automatic annotators more and more precise.

Based on the assumption that visually similar images are more likely to share common labels, many models of nearest nonparametric neighbors have been developed.

Current methods of automatic annotation can be used for video sequences and are divided into two categories: learning-based methods and search-based methods. The automatic methods of annotation of images based on learning generally build a statistical model for the joint distribution of the components based on the image annotations and their visual characteristics, these methods can be divided into methods based on supervised learning and unsupervised learning.

Generally, accuracy of unsupervised methods is lower than the supervised methods [5].

The task of automatically generating the description of an image involves, on the one hand, the understanding of the image, that is, the detection and recognition of the important objects of the scene, as well as their attributes and the relationships between them, and On the other hand, the generation of sentences that synthesize such information in semantically and syntactically correct way. The understanding of the image by the computer requires the extraction of features, for which there are multiple techniques, which can be classified into two groups: 1) Traditional techniques of *machine learning* and 2) Techniques based on *deep learning*.

Traditional machine learning techniques, such as Local Binary Patterns (LBP), Scale-Invariant Feature Transform (SIFT) and Histogram Oriented Gradients (HOG), allow the extraction of characteristics that are subsequently processed by a classifier, such as Support Vector Machines (SVM), however, this approach is based on obtaining very specific characteristics, so it is not possible to generalize the method for heterogeneous input data [8].

Among the methods of automatic annotation that employ traditional techniques of machine learning, are the so-called template-based and those based on recovery. The former use templates with blank spaces that fill up as certain objects, attributes or actions are detected. These templates can be fixed or based on language models and are able to generate grammatically correct annotations. On the other hand, recovery-based approaches obtain the annotations of a set of existing annotations by comparing the images to determine a series of candidate annotations and then select the best of them. The annotations generated by this method are syntactically correct but have semantic errors and are not very specific.

On the other hand, *deep learning* techniques are able to automatically determine which are the relevant characteristics for the classification of heterogeneous input data, provided that the training set is sufficiently large. In general, *convolutional neural networks* (CNN) followed by a *Softmax* classifier, are used in the last layer to

obtains the corresponding score for each class, and a recurrent neural network *neural network* to generate the annotations [8].

The methods of annotation by means of *deep learning* can be classified according to the type of mapped characteristics: on the one hand there are the methods based on visual space, which extract the characteristics of the image and then pass them through a decoder of language to generate the labels. On the other hand, methods based on multimodal space learn such space of the image-annotation set and pass this complete representation by the language decoder.

Regarding the type of learning, the most common method is based on supervised learning. The methods of supervised learning can be classified into: 1) architecture encoder-decoder, 2) compositional architecture, 3) based on attention, 4) based on semantic concepts, 5) stylized labels, 6) new based on objects and 7) dense labeling [8].

Other types of learning used for automatic annotation are reinforcement learning, where an agent generates the labels through an exploration and rewards system, and unsupervised learning, through the generative adversarial networks (GANs).

Regarding the number of annotations, the methods can be divided into those that generate dense annotations, that is, they provide annotations for each region of interest in the scene, while the complete scene annotation methods generate only one label for the whole scene.

Another type of classification is according to the type of architecture: encoder-decoder architecture or compositional architecture. In the encoder-decoder architecture, the characteristics of the image are extracted by means of a CNN network and used as inputs of a language model to generate the annotations. On the other hand, annotation methods based on compositional architecture consist of a set of independent functional blocks; The first block is a CNN network that extracts the semantic concepts from the image.

Once the characteristics of the image have been obtained, multiple candidate labels are generated with a language model and finally, these are ranked by means of a deep multimodal similarity model and the best qualified is selected [3].

## 3 Metrics for Evaluation

To evaluate the results of automatic annotation models and perform comparisons, various metrics are used. The most reliable, but inefficient, is that a group of human evaluators assign a score to each label with a scale of 1 to 4, where 4 corresponds to an image described without any error, 3 to a description with minimal errors, 2 to a description somewhat related to the image and 1 to a description without any relation to the image. Two evaluations of different evaluators are obtained for each annotation and the typical percentage of agreement among evaluators is observed.

On the other hand, there are automatic evaluation metrics that measure the desirable properties in the annotation of an image, such as grammar, coherence, salience, that is, describing the main aspects of the image, the veracity, fidelity in the description and consensus with the descriptions of human experts, among others.

Among the metrics that can be calculated automatically if the groundtruth is available, the most popular is BLEU, which estimates the accuracy between the automatically generated n-grams and the reference n-grams.

BLEU evaluates in a range of 0 to 1 the annotations, where 1 equals an annotation identical to the original reference, so even a human will hardly get this qualification [14]. Other useful metrics are perplexity, ROUGE, METEOR and CIDEr.

ROUGE is a method for the automatic evaluation of abstracts, based on n-grams. For the evaluation of the method, the DUC 2001, 2002 and 2003 databases were used, which include summaries of around 100 words of simple documents, short summaries of around 10 words of simple documents and multi-document summaries of around 10 words. [11].

METEOR is a method initially conceived for the automatic evaluation of translations based on unigrams.

The evaluation of the method was carried out with the databases DARPA/TIDES 2003 Arabic-to-English and Chinese-to-English obtaining a greater correlation with human judgment than BLEU [1, 10].

The method called *Consensus-based Image Description Evaluation*, or CIDEr [18] by its acronym in English, proposed in 2015, is a paradigm for the evaluation of image descriptors by means of human consensus and arises due to that other popular methods, such as BLEU and ROUGE, have shown little correlation with human judgment, that is, how similar an automatic description is with one made by humans.

CIDEr measures the similarity between automatically generated sentences and sentences written by humans, so it presents a high consensus with human judgment and inherently evaluates grammar, salience and precision.

## 4 Databases

The databases used for the evaluation of automatic annotation methods are composed of images and their corresponding descriptors. The most popular are Pascal VOC, Flickr8k, Flickr30k and MS COCO.

The project called Pascal VOC (Visual Object Classes), which took place from 2005 to 2012 and was a reference among the scientific community of computer vision dedicated to the recognition of objects through machine learning, maintains a server through which, provides databases for the recognition of object classes, as well as access to different standardized comparison and evaluation methods.

The images of the database come from the Flickr Web and were selected by maximizing the variability of size, orientation, pose, lighting, location and occlusion conditions of each object and minimizing the bias towards desirable conditions, such as images of centered objects and good lighting [6]. The database consists of 11540 images of 20 different classes, divided into two subsets, the training and the evaluation and each instance in an image is annotated with the class to which it belongs, a polygon that surrounds it and attributes such such as orientation, truncation and difficulty.

Flickr 8K and Flickr 30K are databases of just over 8,000 and 30,000 images respectively, which were obtained from the Flickr web and manually annotated using the services of Amazon's Mechanical Turk. Each annotator received a payment of $ 0.1 for annotating five images, so the total cost of annotating the Flickr 8K database was around $812.00 and it took an average of 3 minutes for each person to make the 5 annotations. The images of Flickr 8K and Flickr 30K were chosen one by one in order to show people and animal actions, rather than simple scenarios and each annotation is a complete descriptive statement of the image [4, 15].

The MS COCO database was developed by Microsoft and is aimed at the task of object recognition. The images that comprise it show common objects that are in their normal context and everyday scenes composed of a large number of elements. There are 91 classes and each instance in an image is surrounded by a polygon and labeled. In total, the database has 328 thousand images and 2.5 million instances labeled [12].

Most of the databases used in the evaluation of automatic annotations have about 5 descriptions per image, however, some studies argue that such a small number of sentences is insufficient to determine how most humans would describe the image, for which the PASCAL-50S and ABSTRACT-50S databases containing 50 descriptions per image [18] have been created, however they contain fewer images.

## 5 Methods for Automatic Image Annotation

Since the annotation of is a key step towards the images retrieval and is a complex and non-scalable task, there are a large number of commercial and non-commercial methods with increasingly performance and characteristics. Among the commercial systems of better performance for the automatic annotation of images, highlight *Microsoft azure* and *Show and tell*, proposed by *Google*.

Although many previous works have given solution to the problem of annotation of images separating it into two subproblems (the knowledge

of the elements of the image, as well as their attributes and actions, and their expression in natural language) and resolving them independently, *Show and tell* of *Google* [19] proposes to analyze them jointly taking an image as input and training a network with the purpose of maximizing the probability of obtaining a sequence of words that describes the image in the best possible way.

*Show and tell* proposes a deep convolutional neuronal network for the extraction of characteristics and the coding of the image, because this type of networks are able to obtain a representation synthesized in a vector of fixed length. Subsequently, the last hidden layer of the CNN network is used as input to an RNN decoder that generates the text descriptors.

For its part, *Microsoft* [16] presents a method focused on the generation of quality labels outside the training domain as part of the *Microsoft cognitive services*. Additionally, the method is able to identify key entities, such as celebrities or reference sites.

Likewise, *Microsoft* discusses the relevance of state of the art results, obtained from the analysis of images of controlled environments with a distribution similar to the examples of training data. Additionally, the reliability of metrics such as BLEUR, METEOR and CIDEr is refuted due to its discrepancy with human judgment.

Additionally, a confidence model is proposed to assign a score to the annotations obtained for difficult cases, and to evaluate the quality of the annotations generated, a series of human evaluations is carried out. The reported results improve the performance of other state-of-the-art methods, both in databases of controlled domains, such as MSCOCO, as well as free domain databases, such as Adobe-MIT FiveK and random images retrieved from Instagram.

The model proposed by Microsoft consists of four independent components and trained separately. The first one is a network-based vision model that detects a broad spectrum of visual concepts. The second, a language model for the generation of candidate labels and a deep multimodal semantic model to evaluate said candidates.

The third, a recognition model of entities to identify celebrities and popular sites, and fourth, a classifier to weigh the level of reliability of each label.

For the generation of annotations, a maximum entropy language model (MELM) and a deep multimodal similarity model (DMSM) are used. These two networks are trained together with a database formed by image-annotation pairs.

Among the most outstanding results presented by Microsoft, it was found that in the data coming from Instagram, that is, free domain, the system reached a satisfaction score of 49.5 %, because many of the images contained filters or were abstract figures.

## 6 Implementation of Non-Commercial Open Source Model *Let me see*

There are some open source image annotators that provide functions and methods that allow to speed up the implementation process, whether for the execution of tests, making changes or comparing results. Among them, *Let me see* [2] is implemented through neural networks and uses an encoder-decoder architecture that solves the problem of image annotation analogously to a language translator.

The automatic annotator *Let me see* uses the encoder to read the input image and encode it into a fixed-length vector, and uses the decoder to read the encoded image and generate the output annotation. For the integration of the encoder and decoder stages, a merge model is used, which combines the encoded form of the image and the encoded form of the annotations as input to the decoder that generates the output annotation.

*Let me see* is a free-code automatic annotator and is based on another remarkable performance automatic annotator, called *Show and tell* and developed by *Google*. Among the code that is provided, a preprocessor is included for the *Flickr8k* database that is suggested for training. For the extraction of image characteristics, the model uses a convolutional neuronal network and for the generation of the sentences, a recurrent neural network.

Additionally, for the implementation of *Let me see* it is proposed the use of a pre-trained model for the stage of the convolutional network, *VGG16*, which is available as part of the *Keras* library, in Python. A fully-connected layer with a SoftMax activation function is used at the output.

For the generation of the model, the authors point out the need to optimize the value of the most relevant parameters, which were determined empirically based on their high impact on the BLEU and ROUGE evaluation metrics. These parameters are the learning rate, the dropout, the size of the outputs of the LTSM layer and the number of epochs, and their proposed values are:

$$\begin{aligned} \text{Learning rate} &= 0.00051 \\ \text{Dropout} &= 0.35 \\ \text{Size of the outputs of LTSM} &= 400 \\ \text{Epochs} &= 14 \end{aligned}$$

To establish the optimal number of epochs in 14, a maximum value of 20 was established, obtained empirically and a stop condition available in the *Keras* library.

The evaluation of the results was performed qualitatively and quantitatively. For the qualitative evaluation, four categories were established for human evaluators to classify the semantic concordance between image-annotation pairs. The four possible categories are: 1) description without errors, 2) description with minor errors, 3) there is something related to the image and 4) there is nothing related to the image. Of the results of the qualitative evaluation, only examples are presented and there is no statistical analysis.

Since the set of tests consists of images from the Flickr8k database, the groundtruth was available and it was possible to carry out the quantitative evaluation of the performance by means of the BLEU and ROUGE metrics and make the comparison with other methods of the state of art, however, in this research we propose the analysis of images obtained from a video surveillance system for which there is no groundtruth, so it will not be possible to make a quantitative comparison of performance.

Since the authors do not provide the model with which they performed the tests that are reported, the first step was to generate it using the same parameters. The following is the methodological design followed for the realization of tests in video sequences of real images and the results obtained from the annotation. The validation of the performance is performed qualitatively, the results are discussed and future work is proposed.

**Fig. 1.** Empty parking lot

## 7 Annotation of Video Surveillance Sequences

In order to analyze the possible applications of automatic annotation in real video surveillance sequences, the capture of videos with an approximate duration of 20 minutes in the parking lot of a public university was carried out. To select the environment where the tests were conducted, the performance of the annotator was also analyzed inside a classroom, however the best results were obtained outdoors.

Four video-surveillance sequences were obtained and each of the frames was annotated. The videos consist of 30 frames per second, so for each one, around 36,000 images were processed. As a result, a repository of structured text strings was generated in the following way: *"frameID_Annotation"*.

Although it would be extremely complicated to perform the evaluation of the results of the automatic annotator by humans for the amount of images that have been processed, the problem is simplified because in general, few changes are observed from one frame to another.

**Fig. 2.** Pedestrian crossing to the side of the parking lot

The task of automatic annotation of frames of video surveillance sequences allows analyzing the sensitivity of the annotator with respect to minimum changes in the scenes. The video surveillance sequences that have been used can be characterized under the following parameters:

1. The images are taken outdoors.
2. The images have been obtained during the day.
3. The environment has not been controlled.
4. The lighting conditions vary.
5. The image is relatively fixed most of the time except for the occurrence of a few events.
6. Events can happen in any plane (first, second plane).
7. The events that happen in the foreground are more relevant than those that happen in the background.
8. Typical events are pedestrian crossing, passing cars, passing bicycles, passing animals or movements caused by the wind.
9. The more relevant an event is, in general, its duration will be shorter. A pedestrian takes about 3 seconds to cross in the foreground and around 10 seconds, to cross in the background.
10. A single fixed camera was used to obtain the video sequences.

To simplify the analysis and evaluation of image-annotation couples, three basic considerations are taken into account:

1. Most of the time, the image is fixed, except for variations in lighting or movements caused by the wind.
2. A few relevant events may occur, and their average total duration is less than 10% of the total time of the video sequence.
3. Due to the characteristics of the database used for training, the relevant events are reduced to the presence of humans, that is, it is not possible to detect the entry or exit of cars in the parking lot.

Taking these considerations into account, it is assumed that approximately 90% of annotations in the repository should be the same or very similar, however, it would still be a complex task for a human evaluator to validate the remaining 10%, which is equivalent to 14400 couples annotation-frame. However, it is still possible, in theory, to reduce the dimensionality of the problem, because each event also has very similar characteristics among its frames. Fig. 2 shows the example of the event in which a pedestrian crosses to the side of the parking lot. This event lasts approximately 9 seconds, which is equivalent to about 3000 frames, which would be described similarly by a human being.

However, these conditions are not met in practice. Minimal variations in lighting conditions or movements caused by the wind greatly affect the interpretation of the image and result in different annotations in almost equal frames, so it is not feasible to carry out the manual validation, by a human being in a reasonable time, of all the frames of the video surveillance sequences used in this analysis.

The qualitative evaluation of 500 frames was performed, which is equivalent to 16.6 seconds of video, period during which the event of a pedestrian crossing to one side of the parking lot occurred, with a duration of 9 seconds.

The evaluation was carried out taking into account the categories 1) description without

errors, 2) description with minor errors, 3) there is something related to the image and 4) there is nothing related to the image. Fig. 1 and Fig. 2 correspond to the two different scenes observed during the proposed time period, the empty parking lot and the crossing of a pedestrian.

The annotations obtained for frames with minimal variations to those of Fig. 1 and 2, only with changes in lighting or movements caused by the wind, as well as their frequency of appearance, are summarized in Table 1.

The first case, where the video surveillance sequence captures only the empty parking lot, consists of 224 frames. Some examples of valid annotations, made by humans beings, could be:

— Outdoor parking lot next to a park.

— Seven cars parked in a parking lot.

— Green area next to a small parking lot.

— Cars in parking lot on the street.

— Garden full of trees with parking lot.

As seen in Table 1, the annotations obtained automatically are very different from the actual content of the image and the descriptions that a human could make. We could classify them as ”there is something related to the image” only because it has been detected that it is an image outside or on the street, however in all cases the presence of one or two people is detected, which are not part of the image.

For the second case, of the frames in which a pedestrian crosses one side of the parking lot, a human being could perform the following annotations:

— A man walking next to a parking lot.

— Man dressed in black walks down the street.

— A man walks on the sidewalk next to a parking lot.

— A man holding a folder is going to cross the street.

— A man walks beside the parking lot of a park.

**Fig. 3.** Partial pedestrian occlusion

For this event, as can be seen in table 1, the annotations continue to detect some elements that are part of the scene, such as the man in the black shirt, but they do not succeed in the activity he is doing.

It is necessary to consider that some variations and errors in the annotations may be associated with occlusions, as in the case of the frame shown in Fig. 3.

## 8 Conclusions

The implementation of an open-source automatic image annotator was performed and tests were carried out with frames of a real video surveillance system placed in the parking lot of a university campus.

For the training of the model, the Flickr8K database was used, which has the characteristic of being made up of images of people or animals performing some action, so in frames where there are no people in the parking lot, the annotations are little or nothing related to visual content. On the contrary, in the frames where people are observed in the parking lot, the results improve and annotations are obtained somewhat related to the visual content of the image.

Since the images of the video surveillance system are very different from those used for training the model, the results are imprecise. It would be possible to improve them by replacing the training database with one generated from the images of the video surveillance system and manually annotated by humans.

**Table 1.** Automatic annotations of video surveillance sequences

| Image | Annotation | Instances | Percentage |
|---|---|---|---|
| Empty parking lot | man in red shirt is sitting on the street | 31 | 13.8% |
| | two people are playing in the background | 169 | 75.5% |
| | two people are playing in the air in the background | 24 | 10.7% |
| | *Total* | 224 | 100% |
| Pedestrian crossing | man in black shirt is sitting on the street | 10 | 3.6% |
| to the side of the parking lot | man in red shirt is sitting on the street | 127 | 46.0% |
| | two people are playing in the background | 78 | 28.3% |
| | two people are playing in the air in the background | 15 | 5.4% |
| | man in red shirt is jumping over the street | 39 | 14.1% |
| | man in red shirt is riding his bike on the street | 7 | 2.6% |
| | *Total* | 276 | 100% |

It would also be possible to validate if the results are more precise by replacing the original database or complementing it with new image-annotation pairs.

It is possible to improve the quality of the proposed analysis, increasing the time when the parking lot is empty so that the proportion is representative, since in this example, the time during which the pedestrian crossed the parking lot is greater than the time the parking lot was empty, but in reality, the time when parking is empty is much longer.

Although an approach based on the automatic generation of annotations from a database of image-annotation pairs, can provide good results in images similar to those of the training set, it is not possible with this method to achieve the quality of the annotations that a human would make in a free domain.

## 9 Future Work

As future work, the quality of the automatic annotations obtained will be improved by enriching the database with specific examples of the parking lot where the tests are being carried out. These new images will be annotated manually and the intrinsic properties of the video, such as the little variation between consecutive frames, will be used for this purpose.

Other automatic image annotator based on different architecture, such as the template model, will also be analyzed. Although this type of architecture does not allow to generate very varied annotations, it is possible that for an analysis of the frames of a fixed camera video surveillance system, the results are better than those obtained with the decoder-decoder model.

Likewise, the analysis will be performed in different scenarios, with which it will be possible to verify the generality of the proposed system and make an accurate estimate of the number of images required for the training, as well as the scalability of the method.

## References

1. **Banerjee, S. & Lavie, A. (2005).** METEOR: An automatic metric for MT evaluation with improved correlation with human judgments. *Proceedings of the acl workshop on intrinsic and extrinsic evaluation measures for machine translation and/or summarization*, pp. 65–72.
2. **Bengoechea Isasa, J. I., .** Let me see: diseño de un generador automático de descripciones de imágenes.
3. **Bernardi, R., Cakici, R., Elliott, D., Erdem, A., Erdem, E., Ikizler-Cinbis, N., Keller, F., Muscat, A., & Plank, B. (2016).** Automatic description generation from images: A survey of models, datasets, and evaluation measures. *Journal of Artificial Intelligence Research*, Vol. 55, pp. 409–442.
4. **Chen, X. & Lawrence Zitnick, C. (2015).** Mind's eye: A recurrent visual representation for image caption generation. *Proceedings of the*

*IEEE conference on computer vision and pattern recognition*, pp. 2422–2431.

5. **Cheng, Q., Zhang, Q., Fu, P., Tu, C., & Li, S. (2018).** A survey and analysis on automatic image annotation. *Pattern Recognition*, Vol. 79, pp. 242–259.
6. **Everingham, M., Van Gool, L., Williams, C. K., Winn, J., & Zisserman, A. (2010).** The Pascal visual object classes (voc) challenge. *International journal of computer vision*, Vol. 88, No. 2, pp. 303–338.
7. **Ghoshal, A., Ircing, P., & Khudanpur, S. (2005).** Hidden Markov models for automatic annotation and content-based retrieval of images and video. *Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pp. 544–551.
8. **Hossain, M. Z., Sohel, F., Shiratuddin, M. F., & Laga, H. (2019).** A comprehensive survey of deep learning for image captioning. *ACM Computing Surveys (CSUR)*, Vol. 51, No. 6, pp. 1–36.
9. **International Standard (2017).** ISO/IEC DIS 15938-15. Compact descriptors for video analysis. Organization for Standardization and International Electrotechnical Commission.
10. **Lavie, A. & Agarwal, A. (2007).** METEOR: An automatic metric for MT evaluation with high levels of correlation with human judgments. *Proceedings of the second workshop on statistical machine translation*, pp. 228–231.
11. **Lin, C.-Y. (2004).** ROUGE: a package for automatic evaluation of summaries. *Workshop on Text Summarization Branches Out, Post-Conference Workshop of ACL 2004, Barcelona, Spain*.
12. **Lin, T.-Y., Maire, M., Belongie, S., Hays, J., Perona, P., Ramanan, D., Dollár, P., & Zitnick, C. L. (2014).** Microsoft coco: Common objects in context. *European conference on computer vision*, Springer, pp. 740–755.
13. **Murthy, V. N., Maji, S., & Manmatha, R. (2015).** Automatic image annotation using deep learning representations. *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*, pp. 603–606.
14. **Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002).** BLEU: a method for automatic evaluation of machine translation. *Proceedings of the 40th annual meeting on association for computational linguistics*, Association for Computational Linguistics, pp. 311–318.
15. **Rashtchian, C., Young, P., Hodosh, M., & Hockenmaier, J. (2010).** Collecting image annotations using Amazon's Mechanical Turk. *Proceedings of the NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk*, Association for Computational Linguistics, pp. 139–147.
16. **Tran, K., He, X., Zhang, L., Sun, J., Carapcea, C., Thrasher, C., Buehler, C., & Sienkiewicz, C. (2016).** Rich image captioning in the wild. *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pp. 49–56.
17. **Uricchio, T., Ballan, L., Seidenari, L., & Del Bimbo, A. (2017).** Automatic image annotation via label transfer in the semantic space. *Pattern Recognition*, Vol. 71, pp. 144–157.
18. **Vedantam, R., Lawrence Zitnick, C., & Parikh, D. (2015).** Cider: Consensus-based image description evaluation. *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 4566–4575.
19. **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.
20. **Zhang, R., Zhang, Z., Li, M., Ma, W.-Y., & Zhang, H.-J. (2005).** A probabilistic semantic model for image annotation and multimodal image retrieval. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, volume 1, IEEE, pp. 846–851.

*Article received on 30/10/2019; accepted on 08/03/2020.*
*Corresponding author is David Pinto.*