International Journal of New Computer Architectures and their Applications (IJNCAA) 8(3): 119-124
The Society of Digital Information and Wireless Communications, 2018 ISSN 2220-9085 (Online); ISSN 2412-3587 (Print)

Content Based Image Retrieval using Multimodal Data based on CCA

Ismail A. El Sayad, Mohammad A. Bazzoun, Hawraa I. Younes, Laila M. Ghoteime, Samih Abdulnabi
Department of Computer and Communication Engineering
Lebanese International University, Beirut, Lebanon
Ismail.sayad@liu.edu.lb ; mohammad.bazzoun@liu.edu.lb ; hawraa.younes@outlook.com ;
laila.ghoteime87@gmail.com ; samih.abdulnabi@liu.edu.lb

***Abstract*— The development of CBIR is based on Multimodal analysis, especially for a set of images associated with some text. Multimodality is used in multimedia to improve data retrieval from the multimedia scope. Visual features and unstructured text annotations processed, analyzed, and resolved within two data modalities simultaneously, which is referred to Multimodality. In this paper, a method is proposed for combining visual and textual data to improve the performance of image retrieval from annotated data set by representing image contents and text semantics in a Multimodal analysis space.**

Keywords—CBIR; Multimodality; CCA; BOW

I. INTRODUCTION

Nowadays, visual information plays an important role in many fields such as medical treatment, crime prevention, hospitals, engineering, journalism...etc [1]. We consume and produce a large amount of visual media with the progression of multimedia technologies and the advances in data storage. Such techniques help us for image capturing, processing, storage, and transmitting, hence enabling users to access data from any place and provide the use of digital images in different fields. However, this will raise the question of which effective methods to access, search and navigate through the data to retrieve the information needed.

Visual information retrieval has attracted great interest too, due to the evolution of image creation tools. In order to deal with this visual information, insisting requests of algorithms that can efficiently meet the needs.

In general, searching images from large database has adopted two different approaches: the first is Text-Based Image Retrieval (TBIR) systems where searching is based on annotated images, and the second is based on image content information (CBIR) which uses visual contents of the images described in the form of low-level features [2].

In TBIR systems, a user provides a query in terms of a keyword and the system will return images similar to the query [3]. The TBIR systems are fast since it applies string matching which needs less computationally time compared to CBIR. But, there are some drawbacks of this type of an image retrieval system: first, a considerable level of human labor is required for manual annotation. Second, the annotation inaccuracy due to the subjectivity of human perception [1]. Additionally, it is sometimes difficult to express in words the visual content of images, and by that decreasing the performance of the keyword-based image search [3].

In order to overcome the limitations enjoined by TBIR systems, CBIR was developed as an alternative. In a typical CBIR system as illustrated in Fig. 1, image low-level features like color, shape, texture and spatial locations are represented in the form of a feature vector. Note that in some CBIR approaches, colored images do not undergo pre-processing, as they are distorted with noise resulting from devices/sensors; thus, to improve the accuracy of retrieval, we may use effective filters to remove this noise. Pre-processing is necessary when results are used for human analysis. There are several color filters available for this purpose [6].

The feature database is formed by the feature vectors of the image. The retrieval process is initiated when a user queries the system using a query image. The query image is converted into a feature vector. The similarity measure is employed to calculate the distance between the feature vectors of the target images in the feature database and the retrieval is performed using an indexing scheme.

In addition to color and texture, spatial location can be considered as a useful factor in region clustering. For instance, ‘sky’ and ‘sea’ share the same color and texture features, yet they have different spatial locations, since sky appears at an image’s top location, while sea appears at its bottom. Spatial locations are defined as upper, bottom or top based on the location of the needed region in the testing image [7]. The search focused on the region centroid and its minimum bounding rectangle to provide the retrieval system with spatial location information, where the spatial center of a region was used to represent the same spatial location [8].

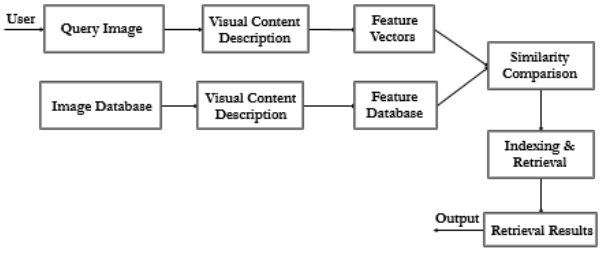

Fig. 1- Typical CBIR System

Whereas there are also some drawbacks for CBIR system where it doesn't support textual queries and it doesn't capture "semantics" by mean that it's possible to answer query 'Red ball' with 'Red Rose' [4]. Thus, retrieval systems tend to incorporate user’s relevance feedback to further improve the

retrieval process and produce more meaningful retrieval results [5] and thus reducing the 'semantic gap' between low-level features and high-level features.

The paper is organized as follows: Section II presents the Multimodality information retrieval methodology based on the proposed work. Section III introduce the proposed method of Image Retrieval utilizing Multimodality. Section IV shows and interprets the result of the proposed approach. Finally, our conclusions are presented in Section VI.

II. Multimodality in Information Retrieval

Human interaction is the major difference between CBIR and TBIR systems. Humans tend to use high-level features (concepts), such as tags, text descriptors, and keywords, to interpret images and measure their similarity [10]. The difference between the limited descriptive power of low-level image features and the richness of user semantics is referred to as the 'semantic gap' [1].

One way of bridging this gap is the use of both visual and textual approach which is introduced to increase the system's performance by using both information. The goal is to join TBIR and CBIR systems into one system known as CBIR system for Multimodal data or Multimodality. Early fusion and late fusion were used to perform this also, but they came up with some drawbacks, where early fusion integrate both data modalities before a user request is received [11].

For the late fusion, it refers to those methods that preserve each data modality separately where. Moreover, late fusion combines the scores of the confidences calculated for the models composed of different features, in such a way that scores represent the possibility of classifying a test sample into the positive class by one specific model [15]. Moreover, late fusion approach lies in its failure to utilize the feature level correlation among modalities [12].

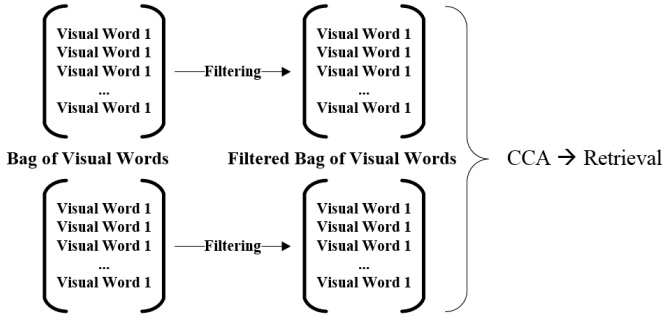

Fig. 2- Multimodal Algorithm using CCA

The approach used in this paper to fuse both visual and textual descriptors together is the Canonical Correlation Analysis (CCA) which is technically able to analyze the data involving multiple sets of variables and is theoretically consistent with that purpose as shown in Fig. 2. The description of the Multimodality algorithm using CCA is presented in the coming sections.

A. Visual Features: Bag-Of-Visual Words

Define Image can be represented by all words that can be generated from it using the so-called visual vocabulary or bag-of-words. The Bag-Of-Visual Words (BOW) [13] is extracted from the whole image, but it's better to form a single bag per each segmented image to avoid mixing between background and foreground.

Researchers are recently using key/local interest points in the retrieval and classification process. Those interest points are called salient image patches that contain rich information about the image. They are then grouped into clusters, such that similar descriptors are assigned to the same cluster. Each cluster is treated as a visual word. When key points are mapped into visual words, we can represent an image as a "bag of visual words", where during the classification task it is used as a feature vector [13]. Once images are represented as bags of visual words, we can classify them by building supervised classifiers. Thus, texture feature becomes a necessary composition in setting up the high-level semantics needed for image retrieval purposes. However, this feature differs from color features as it is not so well-defined, making it not used by some retrieval systems [14].

There are basically 4 steps in order to have a basic implementation of the bag of words object classification:

1. Extracting features from a set of training images.
2. Clustering like-features together into a fixed number of clusters.
3. Constructing histograms of the frequency of features in labeled images containing objects to detect.
4. Evaluating unknown images against the histograms obtained in step 3 to classify objects in the image.

The proposed approach based on the implementation of the bag of words object classification is presented in Section III.

III. Proposed Image Retrieval Approach

The Canonical Correlation Analysis (CCA) is used to improve Image Retrieval using visual and textual data of an image. The visual data (hereby known as Visual Descriptors) are extracted using TOP-SURF [18], whilst the textual data (known as Textual Descriptors) are extracted based on the TF-IDF values of the annotated tags of the images. Using CCA, Visual Descriptors and Textual Descriptors are fused together to produce a multimodal global feature that helps in the retrieval process.

A. Textual Features

Another way to represent the content/semantic of an image is the textual annotations. Textual annotations are texts, tags or keywords representations that mainly describe what is in the image. These annotations might be specific or general, as much specific as they could be, as much as they help in the process of retrieval.

Text annotations are generated in two ways: one way is by manual annotation of images, which requires extra human labor but are very specific to the image, which means higher precision; another way is the use of tagged datasets such as

Flickr or KODAK datasets, which have been already annotated with different tags that could be of high/low importance to the image. Most of the tags annotated to the image, are of low importance and could be noisy, and thus a way to make these tags efficient is by filtering those tags i.e. removing less important tags and keeping tags of high importance (tags that convey the visual content of the image).

After generating tags, TF-IDF values in each image is computed. And thus, each image in the training set and query set can be represented as a vector of textual descriptors.

B. TF-IDF Calculations

In order to represent an image as a multimodal vector representing both visual and textual content, we need to represent the image in two separate vectors: the Visual Descriptors Vector and the Textual Descriptors Vector.

One way to represent these vectors is to use the TF-IDF values of each descriptor. TF-IDF value is the weight of a descriptor in an image with respect to the whole images. TF is the Term Frequency, which represents how frequent this term is in this image.

TF-IDF is an alternative way to evaluate the subject of an object by the words it has. With TF-IDF, words are given weights – TF-IDF measures relevance, not the frequency of occurrence [16]. That is, word counts are interchanged with TF-IDF values through the complete dataset. This measure calculates the number of times that a certain word appears in a specified document, as for words such as "and" or "the" that appear recurrently in all documents, those are analytically reduced.

$$idf\ (\mathrm{t}, \mathrm{D}) = \log \frac{N}{|\{\mathrm{d}\varepsilon\ \mathrm{D} : \mathrm{t}\varepsilon\ \mathrm{D}\}|} \quad \textbf{(1)}$$

IDF is the Inverse Document Frequency, which represents how frequent this term in the whole dataset/document, and it can be calculated in many ways, but the method that we used is the logarithmic IDF method as shown in **(1)**. Where N is the total number of images and is the number of images where the term t appears.

The TF-IDF value is calculated by multiplying the TF value with the IDF value. The method of evaluating TF-IDF is easier to be done manually for textual vectors. As for visual vectors, we use computer-based technique; i.e. TOP-SURF, to extract those descriptors and get their TF-IDF values.

C. Canonical Correlation Analysis

Canonical Correlation Analysis (CCA) [17] is used to combine the visual and textual features, where it applies feature level fusion. Feature fusion is the process of combining two feature vectors to obtain a single feature vector, which is more discriminative than any of the input feature vectors.

CCA is a statistical method to perform multi-view/multi-scale analysis for different data sources. It is a method for correlating linear relationships between multidimensional variables, which is finding projections (linear) for multiple data types that have maximum correlation into a high dimension feature space. CCA integrates Multimodal features to cross-modal features. It projects those features of different data modalities on multiple views into a common subspace to get the maximum correlation between visual and semantic features [12].

Four variables will be needed as inputs to the CCA implementation, two are related to the dataset (Visual & Textual) and the remaining two are related to the query image (Visual & Textual). The first input is "*trainX*" which is an "$n \times p$" matrix containing the first set of training data where n is the number of training samples and p is the dimensionality of the first feature set. Then "*trainY*", an "$n \times q$" matrix containing the second set of training data where q is the dimensionality of the second feature set.

For the query image, "*testX*" is the first input which is an "$m \times p$" matrix containing the first set of test data where m is the number of test samples, the second input is "*testY*", an "$m \times q$" matrix containing the second set of test data.

This implementation gets the train and test data matrices from two modalities X & Y and consolidates them into a single feature set Z and our proposed approach illustrated in the Results section.

IV. Results

The key objective in multimodal data is to enhance the performance of any content-based image retrieval system. To examine clearly the effect of joining two modalities (Visual and Textual), simulations were performed considering visual data alone and another simulation considering the two modalities: Visual and Textual. Our dataset (Training images) composed of 4 different categories: nature, flowers, buses and people, which were tested with several query images. All the images are stored in JPEG format.

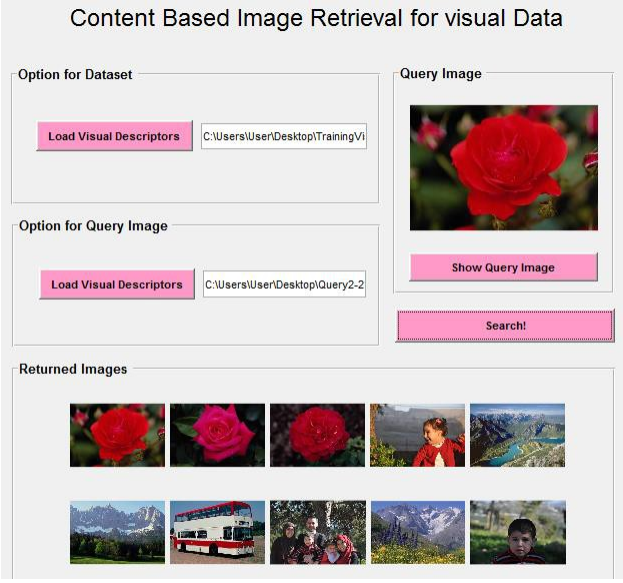

Fig. 3-Result of Visual CBIR system

Precision and Recall shown in **(2)** and **(3)** are used to evaluate the CBIR systems. They are basic measures used in evaluating search strategies. Precision is the fraction of retrieved instances that are relevant. Whilst Recall is the fraction of relevant instances that are retrieved. In the context of Information Retrieval, the precision-recall curve becomes very useful.

$$\text{Precision} = \frac{\text{Number of relevant images retrieved}}{\text{Number of images retrieved}} \times 100 \quad \textbf{(2)}$$

$$\text{Recall} = \frac{\text{Number of relevant images retrieved}}{\text{Number of relevant images in Dataset}} \times 100 \quad \textbf{(3)}$$

$$MAP = \left(\sum_{q=1}^{0} AvgP(q) \right) / Q \quad \textbf{(4)}$$

Another measure shown in **(4)**, is the Mean Average Precision (MAP) for a set of queries which is the mean of the average precision scores for each query. Precision-Recall graphs give more granular detail on how the system is performing.

An example of the retrieval for the visual system for the "Flower" query is illustrated in Fig. 3, where it can easily be seen that the retrieved results are not all accurate; from 10 images we got only 3 relevant images and the rest 7 are not relevant.

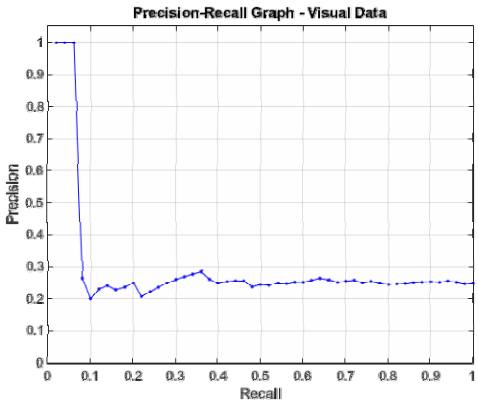

Fig. 4- Precision-Recall of visual CBIR system

We can notice obviously that the Precision value between a Recall ratios of 0.01 to 0.06 is high which is equal to 1. This is because the system is retrieving relevant images for the first three retrievals. This precision decreased to achieve only 0.25 since the system has retrieved an irrelevant image, and it stays low for the last retrieval.

Whilst when applying image retrieval for the "Flower" query in Multimodality using CCA as shown in Fig. 5, from 10 images we got only 9 relevant images and only one irrelevant retrieved image which is an example of our proposed approach that was applied to the whole dataset (Training data) and emanate pretty good results as shown in Fig. 6.

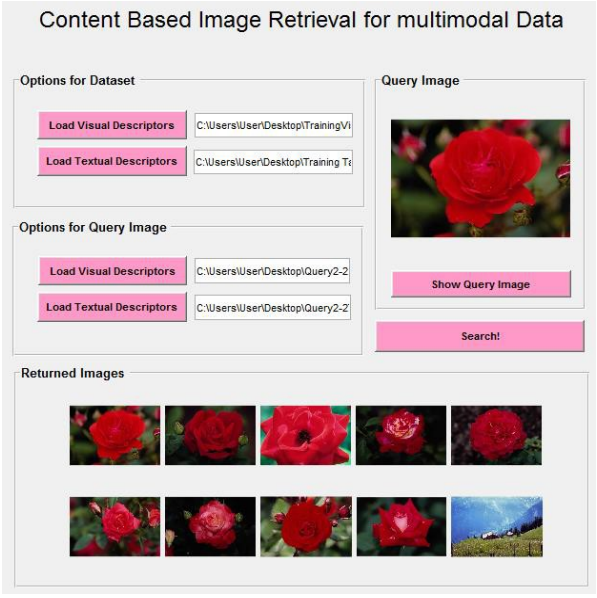

Fig. 5- Result for Multimodal CBIR system

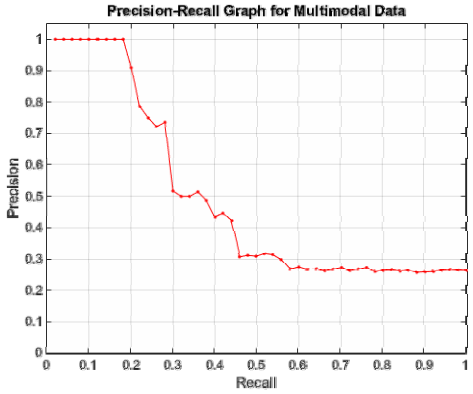

Fig. 6- Precision-Recall of Multimodal CBIR system

The Precision-Recall graph for "Flower" query shown in Fig. 6 present the retrieval of a successive chain of images without touching any of the irrelevant ones for the first 9 retrievals where the Recall ratios are between 0.01 and 0.19, and so the precision will naturally be high. But it will decrease at Recall 0.2 to a value equal to 0.91 when retrieving a non-relevant image. The precision all over the time of retrieval will be affected by the irrelevant results and by that it will still decrease to attempt 0.28 at Recall ratio between 0.58 and 1.

We have managed to retrieve a successive chain of images without changing any of the irrelevant ones for the first 9 retrievals where the Recall ratios are between 0.01 and 0.19, and so the precision will naturally be high. But the precision will decrease, when the Recall becomes 0.2, to a value equal to 0.91 when retrieving a non-relevant image. The precision all over the time of retrieval will be affected by the irrelevant results and by that it will still decrease to attempt 0.28 where the Recall ratio is between 0.58 and 1.

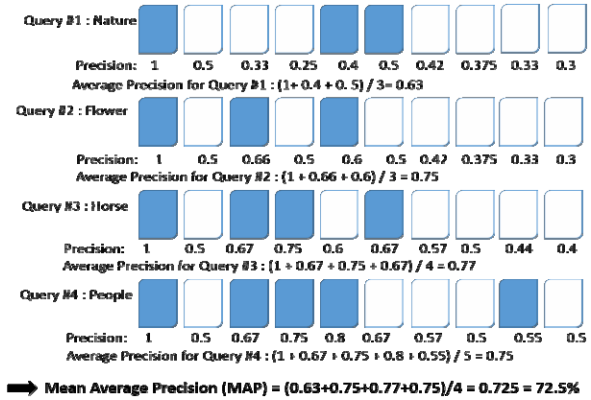

Fig. 7- Mean Average Precision of Visual CBIR system

The mean average precision parameter of the system is calculated for both systems as shown Fig. 8, where the mean average precision which is used to evaluate the efficiency of the system is only 72.6% for the visual CBIR system whereas it is 88.5% for the multimodal CBIR system. Therefore the multimodal CBIR system, which is our proposed approach, present a pretty better efficiency that the CBIR system which may help in reducing time and error while applying image retrieval by decreasing the noise or unrelated retrieved images.

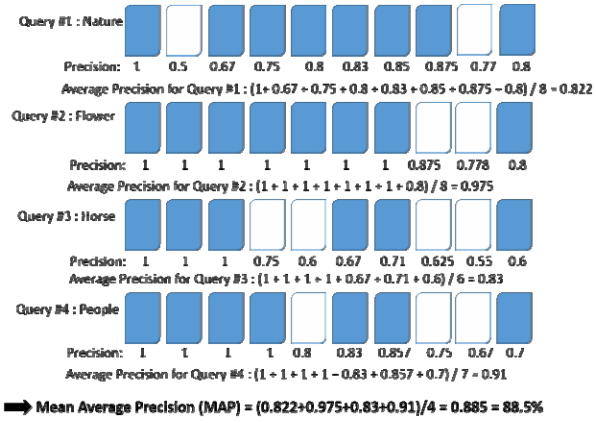

Fig. 8- Mean Average Precision of Multimodal CBIR System

V. Conclusion

The use of multimodality by fusion of visual and textual data became a must in order to improve the performance and efficiency of a content-based image retrieval system. It is an efficient way which makes the accuracy and the precision of the retrieved results very high compared with any system that considers only one modality. Canonical correlation analysis CCA implementation for fusion purpose has demonstrated the high performance of the retrieved results. As for performance measures, the precision-recall graphs were very efficient proofs to confirm the importance and significance of multimodal data.

VI. References

[1] Y. L. . D. Z. . G. L. and W.-Y. M. , "A survey of content-based image retrieval with high-level semantics," Pattern Recognition, vol. 40, no. 1, p. 262–282, January 2007.

[2] S. N. and N. R. , "A NEW CONTENT BASED IMAGE RETRIEVAL SYSTEM USING GMM AND RELEVANCE FEEDBACK," Journal of Computer Science, vol. 10, no. 2, pp. 330-340, 2014.

[3] "Content- based Image Retrieval Approach using Three Features Color, Texture and Shape," International Journal of Computer Applications , vol. 97, p. 0975 – 8887, July 2014.

[4] R. F. L. F.-F. P. P. and A. Z. , "Learning Object Categories from Google's Image Search," in Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on, 17-21 Oct. 2005.

[5] G. W. and A. Y. , "Content Based Image Retrieval Using Enhanced Vocabulary," International Journal of Science and Research (IJSR), vol. 5, no. 5, pp. 2319-7064, May 2016.

[6] A. V. K.N. Plataniotis, "Color Image Processing and Applications," 2000.

[7] . P. K. V. and A. N. , Color Image Processing and Applications, Springer-Verlag Berlin Heidelberg, 2000.

[8] Y. S. . W. W. and A. Z. , "Automatic Annotation and Retrieval of Images," IFIP — The International Federation for Information Processing, vol. 88, pp. 267-280, 2002.

[9] V. M. I. K. and M. S. , "An ontology approach to object-based image retrieval," in Image Processing, 2003. ICIP 2003. Proceedings. 2003 International Conference on, 14-17 Sept. 2003.

[10] M. LEW, N. SEBE, C. DJERABA and R. JAIN, "Content-Based Multimedia Information Retrieval:State of the Art and Challenges," ACM Transactions on Multimedia Computing, Communications and Applications, vol. 2, no. 1, pp. 1-19, February 2006.

[11] M. J. Huiskes and M. S. Lew, "The MIR Flickr Retrieval Evaluation," LIACS Media Lab, Leiden University.

[12] C. J. W. M. R. Z. Y. Z. and X. X. , "Cross-Modal Image Clustering via Canonical Correlation Analysis," in Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence, 2015.

[13] L. K. "Translating images to keywords: problems, applications and progress," in MIS'05 Proceedings of the 11th international conference on Advances in Multimedia Information Systems, Sorrento, Italy, September 19 - 21, 2005.

[14] I. K. S. . I. L. C. and D. S. , "Mining association rules between low-level image features and high-level concepts," in Data Mining and Knowledge Discovery: Theory, Tools, and Technology III, Orlando, April 16, 2001.

[15] J. C. Caicedo , "Multimodal information spaces for content-based image retrieval," in FDIA'09 Proceedings of the Third BCS-IRSG conference on Future Directions in Information Access, Padua, Italy, September 01 - 01, 2009.

[16] "Bag of words - TF-IDF - Deeplearning4j," Skymind. DL4J is licensed Apache 2.0., 2016. [Online]. Available: https://deeplearning4j.org/bagofwords-tf-idf.html#bag-of-words--tf-idf.

[17] . A. Pradeep K., M. A. Hossain, A. El Saddik and M. S. Kankanhalli, "Multimodal fusion for multimedia analysis: a survey," Springer-Verlag, 4 April 2010, p. 345–379.

[18] B. T. E. M. B. and M. S. L. , "TOP-SURF: a visual words toolkit," in Proceedings of the 18th ACM International Conference on Multimedia, Firenze, Italy, 2010.

