Survey on visual sentiment analysis

Alessandro Ortis[1] ✉, Giovanni Maria Farinella[1], Sebastiano Battiato[1]

[1]*Department of Mathematics and Computer Science, University of Catania, Catania, Sicily, Italy*
✉ *E-mail: ortis@dmi.unict.it*

Abstract: Visual Sentiment Analysis aims to understand how images affect people, in terms of evoked emotions. Although this field is rather new, a broad range of techniques have been developed for various data sources and problems, resulting in a large body of research. This paper reviews pertinent publications and tries to present an exhaustive overview of the field. After a description of the task and the related applications, the subject is tackled under different main headings. The paper also describes principles of design of general Visual Sentiment Analysis systems from three main points of view: emotional models, dataset definition, feature design. A formalization of the problem is discussed, considering different levels of granularity, as well as the components that can affect the sentiment toward an image in different ways. To this aim, this paper considers a structured formalization of the problem which is usually used for the analysis of text, and discusses it's suitability in the context of Visual Sentiment Analysis. The paper also includes a description of new challenges, the evaluation from the viewpoint of progress toward more sophisticated systems and related practical applications, as well as a summary of the insights resulting from this study.

1 Introduction and motivations

Nowadays, the amount of public available information encourages the study and development of Sentiment Analysis algorithms that analyse huge amount of users' data with the aim to infer reactions about topics, opinions, trends to understand the mood of the users who produce and share information through the web. The aim of Sentiment Analysis is to extract the attitude of people toward a topic or the intended emotional affect the author wishes to have on the readers. The tasks of this research field are challenging as well as very useful in practise. Sentiment analysis finds several practical applications, since opinions influence many human decisions either in business and social activities. As instance, companies are interested in monitoring people opinions toward their products or services, as well as customers rely on feedbacks of other users to evaluate a product before they purchase it. With the growth of social media (i.e. reviews, forums, blogs and social networks), individuals and organisations are increasingly using public opinions for their decision making [1].

The basic task in Sentiment Analysis is the polarity classification of an input text (e.g. taken from a review, a comment or a social post) in terms of positive, negative or neutral polarity. This analysis can be performed at document, sentence or feature level. The methods of this area are useful to capture public opinion about products, services, marketing, political preferences and social events. For example, the analysis of the activity of Twitter's users can help to predict the popularity of parties or coalitions. The achieved results in Sentiment Analysis within micro-blogging have shown that Twitter posts reasonably reflect the political landscape [2]. Historically, Sentiment Analysis techniques have been developed for the analysis of text [3], whereas limited efforts have been employed to extract (i.e. infer) sentiments from visual contents (e.g., images and videos). Even though the scientific research has already achieved notable results in the field of textual Sentiment Analysis in different contexts (e.g. social network posts analysis, product reviews, political preferences, etc.) [4, 5], the task to understand the mood from a text has several difficulties given by the inherent ambiguity of the various languages (e.g. ironic sentences), cultural factors, linguistic nuances and the difficulty of generalise any text analysis solution to different language vocabularies. The different solutions in the field of text Sentiment Analysis have not yet achieved a level of reliability good enough to

be implemented without enclosing the related context. For example, despite the existence of natural language processing tools for the English language, the same tools cannot be used directly to analyse text written in other languages. Although NLP (Natural Language Processing) and information retrieval researchers proposed several approaches to address the problem of Sentiment Analysis [3], the social media context offers some additional challenges. Beside the huge amounts of available data, typically the textual communications on social networks consist of short and colloquial messages. Moreover, people tend to use also images and videos, in addition to the textual messages, to express their experiences through the most common social platforms. The information contained in such visual contents are not only related to semantic contents such as objects or actions about the acquired picture, but also cues about affect and sentiment conveyed by the depicted scene. Such information is hence useful to understand the emotional impact (i.e. the evoked sentiment) beyond the semantic. For these reasons images and videos have become one of the most popular media by which people express their emotions and share their experiences in the social networks, which have assumed a crucial role in collecting data about people's opinions and feelings. The images shared in social media platforms reflect visual aspects of users' daily activities and interests. Such growing user generated images represent a recent and powerful source of information useful to analyse users' interests. Extracting the emotions inherently underlying images perceived by the viewers could promote new approaches to several application fields such as brand perception, assessment of customer satisfaction, advertising, media analytic, etc. The diffusion of personal mobile devices constantly connected to Internet services (e.g. smartphones) and the growth of social media platforms introduced a new communication paradigm by which people that share multimedia data. In this context, uploading images to a social media platform is the new way by which people share their opinions and experiences. This provides a strong motivation for research on this field, and offers many challenging research problems. The proposed paper aims to introduce this emerging research field. It analyses the related problems, provides a brief overview of current research progress, as well as discusses the major issues and outline the new opportunities and challenges in this area. This paper extends our preliminary work [6], by significantly augmenting the number and type of reviewed papers, as well as the level of details, including

publications that address specific aspects of the issues mentioned in [6] related to features, models, datasets and challenges. This paper further adds a detailed study of the works that tackle specific tasks (e.g. popularity and popularity dynamics, relative attributes, virality, etc.).

Besides the overview of the state-of-the-art, the paper proposes a critical viewpoint on the previous works considering the employed features, datasets and methods. Indeed, Section 2 presents a review of the state-of-the-art works that mostly influenced the research field. In particular, an overview of the most significant works in the field of Visual Sentiment Analysis published between 2010 and 2019 is presented. The literature is presented in a chronological order, highlighting similarities and differences between the different works, with the aim to drive the reader along the evolution of the developed methods. Then, a structured and critical dissertation of the topics and issues that a researcher interested in Visual Sentiment Analysis needs to address is presented in Section 3. These include the choice of the emotional categories, the dataset definition (e.g., existing datasets, sources of data, labelling, size of a good dataset, etc.), as well as considerations about the features that may influence the sentiment toward an image. It provides a complete overview of the system design choices, with the aim to provide a deep debate about each specific issue related to the design of a Visual Sentiment Analysis system: emotions representation schemes, existing datasets and features. Each aspect is discussed, and the different possibilities are compared one each other, with proper references to the state-of-the-art. Section 4 provides a thorough formalisation of the problem, by abstracting all the components that could affect the sentiment associated with an image, including the sentiment holder and the time factor, often ignored by the existing methods. References to the state-of-the-art addressing each component are reported, as well as the different alternative solutions are proposed. The paper also addresses the future perspectives of the field by describing five emerging tasks related to Visual Sentiment Analysis, with proper references, in Section 5. The paper introduces some additional challenges and techniques that could be investigated, proposing suggestions for new methods, features and datasets. In this sense, the proposed paper aims to serve as a guide for research scientist that are interested in Visual Sentiment Analysis. Indeed, this paper is intended to be a reference source of contents in the field, by providing a structured description and analysis of the past works, a detailed presentation of the different existing datasets with the pros and cons, as well as a perspective on current challenges and for future directions. Section 6 reports the insights as outcome of the proposed study. Then, Section 7 concludes the paper by providing a summary of the previous sections and by suggesting directions for future works.

2 State-of-the-art

Visual Sentiment Analysis is a recent research area. Most of the works in this new research field rely on previous studies on emotional semantic image retrieval [7–10], which make connections between low-level image features and emotions with the aim to perform automatic image retrieval and categorisation. These works have been also influenced by empirical studies from psychology and art theory [11–16]. Other research fields close to Visual Sentiment Analysis are those considering the analysis of the image aesthetic [17–21], interestingness [22], affect [23] and popularity [24–27].

The first paper on Visual Sentiment Analysis aims to classify images as 'positive' or 'negative' and dates back on 2010 [28]. In this work, the authors studied the correlations between the sentiment of images and their visual content. They assigned numerical sentiment scores to each picture based on their accompanying text (i.e. meta-data). To this aim, the authors used the SentiWordNet [29] lexicon to extract sentiment score values from the text associated with images. This work revealed that there are strong correlations between sentiment scores extracted from Flickr meta-data (e.g. image title, description and tags provided by the user) and visual features (i.e. SIFT-based bag-of-visual words, and local/global RGB histograms).

In [30] a study on the features useful to the task of affective classification of images is presented. The insights from the experimental observation of emotional responses with respect to colours and art have been exploited to empirically select the image features. To perform the emotional image classification, the authors considered the eight emotional output categories as defined in [31] (i.e. Awe, Anger, Amusement, Contentment, Excitement, Disgust, Sad, and Fear).

In [32] the authors built a large-scale Visual Sentiment Ontology (VSO) of semantic concepts based on psychological theories and web mining (SentiBank). A concept is expressed as an adjective-noun combination called Adjective Noun Pair (ANP) such as 'beautiful flowers' or 'sad eyes'. After building the ontology consisting of 1.200 ANP, they trained a set of 1.200 visual concept detectors which responses can be exploited as a sentiment representation for a given image. Indeed, the 1.200 dimension ANP outputs (i.e. the outputs of the ANP detectors) can be exploited as features to train a sentiment classifier. To perform this work, the authors extracted adjectives and nouns from videos and images tags retrieved from YouTube and Flickr, respectively. These images and videos have been searched using the words corresponding to the 24 emotions defined in the Plutchik Wheel of Emotion [33], a well known psychological model of human emotions. The authors released a large labelled image dataset composed by half million Flickr images regarding to 1.200 ANPs. Results show that the approach based on SentiBank concepts outperforms text based method in tweet sentiment prediction experiments. Furthermore, the authors compared the SentiBank representation with shallow features (colour histogram, GIST, LBP, BoW) to predict the sentiment reflected in images. To this end, they used two different classification models (LinearSVM and Logistic Regression) achieving significant performance improvements when using the SentiBank representation. The proposed mid-level representation has been further evaluated in the emotion classification task considered in [30], obtaining better results.

In 2013, Yuan *et al.* [34] employed scene-based attributes to define mid-level features, and built a binary sentiment classifier on top of them. Furthermore, their experiments demonstrated that adding a facial expression recognition step helps the sentiment prediction task when applied to images with faces.

In 2014, Yang *et al.* [35] proposed a Sentiment Analysis approach based on a graphical model which is used to represent the connections between visual features and friends interactions (i.e. comments) related to the shared images. The exploited visual features include saturation, saturation contrast, bright contrast, cool colour ratio, figure-ground colour difference, figure-ground area difference, background texture complexity, and foreground texture complexity. In this work, the authors considered the Ekman's emotion model [36].

Chen *et al.* [37] introduced a CNN (convolutional neural network)-based approach, also known as 'SentiBank 2.0' or 'DeepSentiBank'. They performed a fine-tuning training on a CNN model previously trained for the task of object classification to classify images in one of a 2.096 ANP category (obtained by extending the previous SentiBank ontology [32]). This approach significantly improved the ANP detection with respect to [32]. Similarly to [32], this approach provides a sentiment feature (i.e. a representation) of an image that can be exploited by further systems.

In contrast to the common task of infer the affective concepts intended by the media content publisher (i.e. by analysing the text associated with the image by the publisher), the method proposed in [38] tries to predict what concepts will be evoked to the image viewers.

In [39] a pre-trained CNN is used as a provider of high-level attribute descriptors in order to train two sentiment classifiers based on Logistic Regression. Two types of activations are used as visual features, namely the fc7 and fc8 features (i.e. the activations of the seventh and eighth fully connected layers of the CNN, respectively). The authors propose a fine-grained sentiment categorisation, classifying the polarity of a given image through a 5-scale labelling scheme: *'strong negative', 'weak negative',*
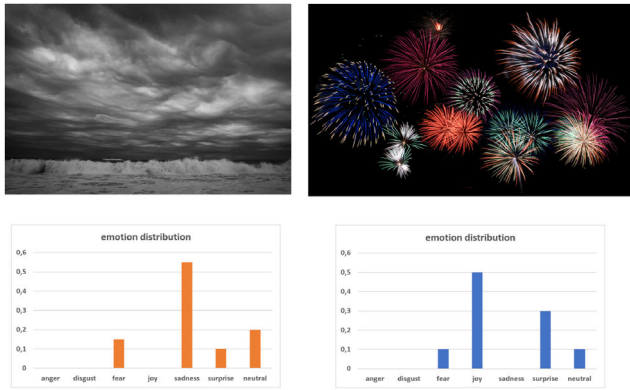
Fig. 1 *Examples of image emotion distributions*

'neutral', 'weak positive', and 'strong positive'. The evaluation of this approach considers two baseline methods taken from the state-of-the-art, namely low-level visual features and SentiBank, both introduced in [32], in comparison with their approaches (the fc7 and fc8 based classifiers). The experimental setting evaluates all the considered methods on two real-world datasets related to Twitter and Tumblr, whose images have been manually labelled considering the above described 5-scale score scheme. The results suggested that the methods proposed in [39] outperform the baseline methods in visual sentiment prediction.

The authors of [40] proposed to use a progressive approach for training a CNN (called Progressive CNN or PCNN) in order to perform visual Sentiment Analysis in terms of 'positive' or 'negative' polarity. They first trained a CNN architecture with a dataset of half million Flickr images introduced in [32]. At the training time, the method selects a subset of training images which achieve high prediction scores. Then, this subset is used to further fine-tune the obtained CNN. In the architecture design they considered a last fully connected layer with 24 neurones. This design decision has been taken with the aim to let the CNN learn the response of the 24 Plutchik's emotions [33]. An implementation of the architecture proposed in [40] is publicly available. The results of experiments performed on a set of manually labelled Twitter images show that the progressive CNN approach obtains better results with respect to other previous algorithms, such as [32, 34].

Considering that the emotional response of a person viewing an image may include multiple emotions, the authors of [41] aimed to predict a distribution representation of the emotions rather than a single dominant emotion from (see Fig. 1). The authors compared three methods to predict such emotion distributions: a Support Vector Regressor (based on hand crafted features related to edge, colour, texture, shape and saliency), a CNN for both classification and regression. They also proposed a method to change the evoked emotion distribution of an image by editing its texture and colours. Given a source image and a target one, the proposed method transforms the colour tone and textures of the source image to those of the target one. The result is that the edited image evokes emotions closer to the target image than the original one. This approach has been quantitatively evaluated by using four similarity measures between distributions. For the experiments, the authors consider a set of seven emotion categories, corresponding to the six basic emotions defined by Ekman *et al.* in [36] and the neutral emotion. Furthermore, the authors proposed a sentiment database called Emotion6. The experiments on evoked emotion transfer suggest that holistic features such as the colour tone can influence the evoked emotion, albeit the emotion related to images with high-level semantics are difficult to be shaped according to an arbitrary target image.

In [42] the textual data, such as comments and captions, related to the images are considered as contextual information. Differently from the previous approaches, which exploit low-level features [23], mid-level features [32, 34] and deep learning architectures [40, 41], the framework in [42] implements an unsupervised approach (USEA - Unsupervised SEntiment Analysis). In [43] a CNN pre-trained for the task of object classification is fine-tuned

to accomplish the task of visual sentiment prediction. Then, with the aim to understand the contribution of each CNN layer for the task, the authors performed an exhaustive layer-per-layer analysis of the fine-tuned model. Indeed, the traditional approach consists in initialising the weights obtained by training a CNN for a specific task, and replacing the last layer with a new one containing a number of units corresponding to the number of classes of the new target dataset. The experiments performed in this paper explored the possibility to use each layer as a feature extractor and training individual classifiers. This layer-by-layer study allows measuring the performance of the different layers, which are useful to understand how the layers affect the whole CNN performances. Based on the layer-by-layer analysis, the authors proposed several CNN architectures obtained by either removing or adding layers from the original CNN.

Even though the conceptual meaning of an image is the same for all cultures, each culture may have a different sentimental expression of a given concept. Motivated by this observation, Jou *et al.* [44] extended the ANP ontology defined in [32] for a multi-lingual context. Specifically, the method provides a multi-lingual sentiment driven visual concept detector in 12 languages. The resulting Multi-lingual Visual Sentiment Ontology (MVSO) provides a rich information source for the analysis of cultural connections and the study of the visual sentiment across languages.

Starting with the fact that either global features and dominant objects bring massive sentiment cues, Sun *et al.* [45] proposed an algorithm that extracts and combines features from either the whole image and 'salient' regions. These regions have been selected by considering proper objectness and sentiment scores aimed to discover affective local regions. The proposed method obtained good results compared with [32, 40] on three widely used datasets presented in [32, 40].

In [46] Katsurai and Satoh exploited visual, textual and sentiment features to build a latent embedding space where the correlation between the projected features from different views is maximised. This work implements the CCA (Canonical Correlation Analysis) technique to build a 3-view embedding which provides a tool to encode inputs from different sources (i.e. a text and an image with similar meaning/sentiment are projected nearby in the embedding space) and a method to obtain a sentiment representation of images (by simply projecting an input feature to the latent embedding space). This representation is exploited to train a linear SVM classifier to infer positive or negative polarity. The authors used a composition of RGB histograms, GIST, SIFT based Bag-of-Words and two mid-level features defined in [32, 47] as visual features. The textual feature is obtained using a Bag-of-Words approach form text associated with the image, crawled from Flickr and Instagram. The sentiment features are obtained starting from the input text and exploiting an external knowledge base, called SentiWordNet [29], a well-known lexical resource used in opinion mining to assign sentiment scores to words.

The works in [48, 49] perform emotional image classification of images considering multiple emotional labels. As previously proposed in 2015 by Peng *et al.* [41], instead of training a model to predict only one sentiment label, the authors considered a distribution over a set of pre-defined emotional labels. To this aim, they proposed a multi-task system which optimises the classification and the distribution prediction simultaneously. In [48] the authors proposed two Conditional Probability Neural Networks (CPNN), called Binary CPNN (BCPNN) and Augmented CPNN (ACPNN). A CPNN is a neural network with one hidden layer which takes either features and labels as input and outputs the label distribution. Indeed, the aim of a CPNN is to predict the probability distribution over a set of considered labels. The authors of [49] changed the dimension of the last layer of a CNN pre-trained for object classification in order to extract a probability distribution with respect to the considered emotional labels, and replaced the original loss layer with a function that integrates the classification loss and sentiment distribution loss through a weighted combination. Then the modified CNN has been fine-tuned to predict sentiment distributions. Since the majority of the existing datasets are built to assign a single emotion ground truth to each image, the authors of [49] proposed two approaches to

convert the single labels to emotional distribution vectors, which elements represent the degree to which each emotion category is associated with the considered image. This is obtained considering the similarities between the pairwise emotion categories [33]. The experimental results show that the approach proposed in [49] outperforms 11 baseline Label Distribution Learning (LDL) methods, including BCPNN and ACPNN proposed in [48].

In [50] the authors extended their previous work [43] in which they first trained a CNN for sentiment analysis and then empirically studied the contribute of each layer. In particular, they used the activations in each layer to train different linear classifiers. In this work, the authors also studied the effect of weight initialisation for fine-tuning by changing the task (i.e. the output domain) for which the fine-tuned CNN has been originally trained. Then, the authors propose an improved CNN architecture based on the experimental results and observations. Then, the authors propose an improved CNN architecture based on the empirical insights.

The authors of [51] exploited a large tweets dataset named Twitter for Sentiment Analysis (T4SA) to fine-tune existing CNN previously trained for objects and places classification (VGG19 [52] and HybridNet [53]) for the task of Visual Sentiment Analysis, where the image labels have been defined based on the tweet texts. The proposed system has been compared with the CNN and PCNN presented in [40], DeepSentiBank [37] and MVSO [44] obtaining better results on the built dataset.

The system proposed in [54] represents the sentiment of an image by extracting a set of ANPs describing the image. Then, the weighted sum of the extracted textual sentiment values is computed, by using the related ANP responses as weights. The approach proposed in this paper takes the advantage of the sentiment of the text composing the ANPs extracted from images, instead of only considering the ANP responses defined in SentiBank [32] as mid-level representations. In particular, the sentiment value of an extracted ANP is defined by summing the sentiment scores defined in SentiWordNet [29] and SentiStrength [55] for the pair of adjective and the noun words of th ANP. A logistic regressor is used to infer the sentiment orientation by exploiting the scores extracted from the textual information, and a logistic classifier is trained for polarity prediction by exploiting the traditional ANP responses as representations. Then, the two schemes are combined by employing a late fusion approach. The authors compared their method with respect to three baselines: a logistic regression model based on the SentiBank mid-level representation, the CNN and PCNN methods proposed in [40]. Experiments show that the proposed late fusion method outperforms the method based only on the mid-level representation defined in SentiBank, demonstrating the contribute given by the sentiment coefficients associated with the text composing the extracted ANPs. However, the CNN and PCNN approaches proposed in [40] exhibit better performances than the late fusion method. The problem of the noise in the text associated with social images by users is very common in the state-of-the-art approaches.

The work in [56] presents a method for Visual Sentiment Analysis that integrates saliency detection into visual attention learning aimed to let the sentiment classifier focus on salient objects in the image. In particular, a fully-connected neural network is trained to extract salient regions in the input image using multi-scale feature maps. During prediction, the image weighted by the attention map is fed to a CNN trained for Visual Sentiment Analysis. The proposed approach has been evaluated on the tasks of polarity prediction and emotion classification using Mikels' 8 emotional model [31].

Ortis et al. in [57] addressed the challenge of image sentiment polarity prediction by proposing a novel source of text for this task, dealing with the issue related to the use of text associated with images provided by users, which is commonly used in most of the previous works and is often noisy due its subjective nature. In particular, the authors built an image sentiment classifier that exploits a representation based on both visual and textual features extracted from an image, and evaluated the performances obtained by using the text provided by users (subjective) and the text extracted from the visual content by using four deep models trained on different tasks. Several experiments have been performed by combining different features based on subjective (i.e. user-provided) and objective (i.e. automatically extracted) text. The experiments revealed that employing a source of text automatically extracted from the images, in lieu of the text provided by users, improves the classifier performances.

In [58] the authors combined the visual and textual information of social media posts to predict the overall sentiment of daily news. The system has been evaluated on a dataset of images crawled from Instagram.

The authors of [59] proposed a joint visual–textual sentiment analysis system trying to exploit more than one modality. However, the authors further incorporated a cross-modality attention mechanism and semantic embedding learning based on bidirectional recurrent neural networks with the aim to design a model able to focus on the visual and textual features that mostly contribute to the sentiment classification. The intuition beyond this approach is that not both text and image contribute equally to the sentiment classification, as visual and textual information differ in their contribution to sentiment analysis. The attention mechanism allows to properly weight input that may deliver inconsistent sentiment, such as the noisy text provided by users [57].

Huang et al. [60] proposed an image–text sentiment model named Deep Multimodal Attentive Fusion (DMAF). This approach defines two attention models aimed to learn effective sentiment classifiers for visual and textual inputs, named Visual Attention Model and Semantic Attention Model, respectively. Furthermore, an additional Multimodal Attention Model is employed with the aim to exploit the correlation between visual and textual features. Then, a late fusion approach is used to combine the three attention models. In [61] the authors trained a joint embedding for images and text for sentiment analysis. In particular, they formulated an ANP detection task in images computed on features extracted from a continuous embedding space that encodes inter-concept relationships between different ANPs. In [62] the authors presented a multi-task framework to combine multimodal information (i.e. images and text) when it is available, while being able to handle the cases when a modality is missing. In particular, this approach adds two auxiliary image-based and text-based classifiers to the classic multimodal framework. Experiments empirically demonstrated that such a multi-task learning approach can also improve generalisation by acting as a regularisation mechanism. The work in [63] presents a method that combines visual and textual features by employing an ensemble learning approach. In particular, the authors combined five state-of-the-art classifiers based on visual and textual inputs. The system classifies a givne image using the eight emotions scheme proposed by Mikels et al. [31]. The authors of [64] combined the use of CNNs and saliency detection to develop a system that first performs a prediction of the sentiment of the whole image. Then, if salient regions are detected, performs the sentiment prediction on the sub-images that depict such sub-areas. Finally, the global and local predictions are fused by a weighted sum.

The paper in [65] presents an overview in the field of affective image content analysis with respect to two main tasks, namely affective gap and perception subjectivity.

Soleymani et al. [66] presented a brief overview of the methods adopted to address three different fields: multi-modal spoken reviews and vlogs, sentiment analysis in humans–machines face-to-face interactions and sentiment of images and tags in social media.

The paper in [67] presents an overview on textual sentiment analysis of social networks contents. Then, a focus on multi-modal approaches that combines textual and visual features for visual sentiment analysis is presented.

The works described in this section have led to significant improvements in the field of Visual Sentiment Analysis. However, these works address the problem considering different emotion models, datasets and evaluation methods (see Table 1). So far, researchers formulated this task as a classification problem among a number of polarity levels or emotional categories, but the number and the type of the emotional outputs adopted for the classification are arbitrary. The difference in the adopted emotion categories

Table 1 Summary of the most relevant publications on Visual Sentiment Analysis. While the early methods were mostly based on hand-crafted visual features, more recent approaches exploit textual metadata and features learned directly from the raw image (i.e. CNN-based representations)

Year	Paper	Input	Output
2010	Siersdorfer *et al.* [28]	hand-crafted visual features	sentiment polarity
2010	Machajdik and Hanbury [30]	hand-crafted visual features	emotional classification [31]
2013	Borth *et al.* [32]	ANP output responses [32]	sentiment polarity and emotional classification [31]
2013	Yuan *et al.* [34]	hand-crafted visual features	sentiment polarity
2014	Yang *et al.* [35]	hand-crafted visual features	emotional classification [31]
2014	Chen *et al.* [37]	raw image	ANP annotation [32]
2014	Xu *et al.* [39]	CNN activations	5-scale sentiment score
2015	You *et al.* [40]	raw image	sentiment polarity
2015	Peng *et al.* [41]	hand-crafted visual features	distribution of emotions
2015	Wang *et al.* [42]	textual metadata	sentiment polarity
2015	Campos *et al.* [43]	raw image	sentiment polarity
2016	Sun *et al.* [45]	image salient regions	sentiment polarity
2016	Katsurai and Satoh [46]	hand-crafted visual features and textual metadata	sentiment polarity
2017	Yang *et al.* [48]	raw image	distribution of emotions
2017	Jufeng Yang and She [49]	raw image	distribution of emotions
2017	Campos *et al.* [50]	raw image	sentiment polarity
2017	Vadicamo *et al.* [51]	raw image	sentiment polarity
2018	Li *et al.* [54]	text from the inferred ANPs [32]	sentiment polarity
2018	Song *et al.* [56]	raw image	sentiment polarity and emotional classification [31]
2018	Ortis *et al.* [57]	hand-crafted visual features and text extracted from the image	sentiment polarity
2019	Felicetti *et al.* [58]	raw image and textual metadata	sentiment polarity
2019	Zhu *et al.* [59]	raw image and textual metadata	sentiment polarity
2019	Huang *et al.* [60]	raw image and textual metadata	sentiment polarity
2019	Corchs *et al.* [63]	raw image and textual metadata	emotional classification [31]
2019	Campos *et al.* [61]	raw image and textual metadata	sentiment polarity
2019	Fortin and Chaib-draa [62]	raw image and textual metadata	sentiment polarity
2019	Wu *et al.* [64]	raw image and textual metadata	sentiment polarity
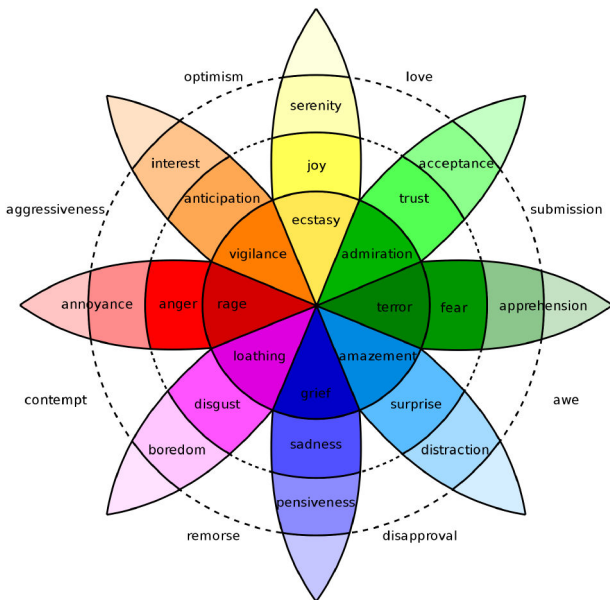
Fig. 2 *Plutchik's wheel of emotions*

makes result comparison difficult. Moreover, there is not a strong agreement in the research community about the use of a universal benchmark dataset. Indeed, several works evaluated their methods on their own datasets. Many of the mentioned works present at least one of the said issues.

Previous surveys give a brief overview of the field followed by a focus on specific issues [65–67]. In contrast, the proposed study introduces the basic task in Visual Sentiment Analysis (i.e.

classifying an image as a positive or a negative) and then extends the discussion by deeply analysing each aspect, covering a larger number of topics, derived tasks and new challenges related to Visual Sentiment Analysis with high level of detail (Fig. 2).

3 Visual sentiment analysis systems

This section provides an overview of the system design choices, with the aim to provide a comprehensive debate about each specific issue, with proper references to the state-of-the-art.

3.1 How to represent the emotions?

Basically, the goal of a Visual Sentiment Analysis system is to determine the sentiment polarity of an input image (i.e. positive or negative). Several works aim to classify the sentiment conveyed by images into two (positive, negative) or three polarity levels (positive, neutral, negative) [28, 32, 40]. However, there are also systems that adopt more than three levels, such as the 5-level sentiment scheme used by Xu *et al.* [39] or the 35 'impression words' used by Hayashi and Hagiwara [68]. Beside the polarity estimation, there are systems that perform the sentiment classification by using a set of emotional categories, according to an established emotion model based on previous psychological studies. However, each emotional category usually corresponds to a positive or negative polarity [30]. Thus, these systems can also be evaluated for the task of polarity estimation. Generally, there are two main approaches for emotion modelling:

- *Dimensional approach:* emotions are represented as points in a 2- or 3-dimensional space (see Figs. 3 and 4). Indeed, as discussed in several studies [11, 13–15], emotions have three basic underlying dimensions: valence, arousal and control (or
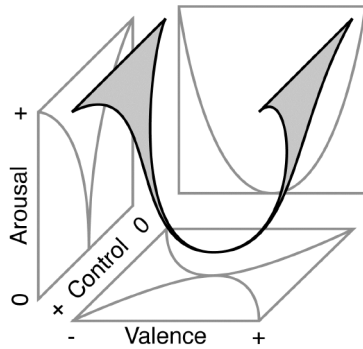
Fig. 3 *Illustration of the 3D emotion space (VAC space), image borrowed from [69]*
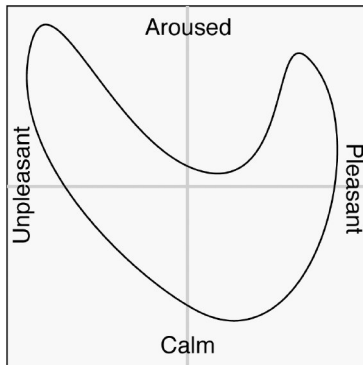
Fig. 4 *Illustration of the 2D emotion space (VA space) obtained considering just the arousal and the valence axis. Image borrowed from [69]*

dominance). However, as can be seen from Fig. 3, the control dimension has a small effect. Therefore, a 2D emotion space is often used, obtained by considering only the arousal and the valence axis.

- *Category approach:* according to this model, there are a number of basic emotions. Having just a few emotional categories are convenient for tasks such as indexing and classification. This set of descriptive words can be assigned to regions of the VAC (Valence-Arousal-Control) space. Thus it can be considered a quantised version of the dimensional approach. The choice of the emotional categories in not an easy task. Since emotion belongs to the psychology domain, the insights and achievements from cognitive science can be beneficial for this problem.

What are the basic emotions? There are several works that aim to ask this question. As we observed in Section 2, the most adopted model is the Plutchnik's Wheel of Emotions [33]. This model defines eight basic emotions with three valences each. Thus, it defines a total of 24 emotions (see Fig. 2).

According to Ekman's theory [36], there are just five basic emotions ('anger', 'fear', 'disgust', 'surprise' and 'sadness'). Another emotions categorisation is the one defined in a psychological study by Mikels *et al.* [31]. In this work, the authors performed an intensive study on the *International Affective Picture System* (IAPS) in order to extract a categorical structure of such dataset [70]. As result, a subset of IAPSs have been categorised in eight distinct emotions: 'amusement', 'awe', 'anger', 'contentment', 'disgust', 'excitement', 'fear' and 'sad'. A deeper list of emotions are described in Shaver *et al.* [71], where emotion concepts are organised in a hierarchical structure. Furthermore, a cognitive perspective analysis on basic emotions and their relations is presented in [72].

As we discussed in this section, there is a wide range of research on identification of basic emotions. By way of conclusion, the 24 emotions model defined in Plutchnik's theory [33] is a well established psychological model of emotions. This model is inspired by chromatics in which emotions are organised along a

wheel scheme where bipolar elements are placed opposite one to each other. Moreover, the three intensities provide a richer set of emotional valences. For these reasons, it can be considered the reference model for the identification of the emotional categories.

3.2 Existing datasets

There are several sources that can be exploited to build Sentiment Analysis datasets. The general procedure to obtain a human labelled set of data is to perform surveys over a large number of people, but in the context of Sentiment Analysis the collection of huge opinion data can be alternatively obtained by exploiting the most common social platforms (Instagram, Flickr, Twitter, Facebook, etc.), as well as websites for collecting business and products reviews (Amazon, Tripadvisor, Ebay, etc.). Indeed, nowadays people are used to express their opinions and share their daily experiences through the Internet Social Platforms.

In the context of Visual Sentiment Analysis, one of the first published dataset is the IAPS [70]. This dataset has been developed with the aim to produce a set of evocative colour images that includes contents from a wide range of semantic categories. This work provides a set of standardised stimuli for the study of human emotional process. The dataset is composed by hundreds of pictures related to several scenes including images of insects, children, portraits, poverty, puppies and diseases, which have been manually rated by humans by means of affective words. This dataset has been used in [30] in combination with other two datasets built by the authors, which are publicly available [www.imageemotion.org].

In [73] the authors considered a subset of IAPS extended with subject annotations to obtain a training set categorised in distinct emotions according to the emotional model described in [31] (see Section 3.1). However, the number of images of this dataset is very low.

In [30], the authors presented the Affective Image Classification Dataset. It consists of two image sets: one containing 228 abstract paintings and the other containing 807 artistic photos. These images have been labelled by using the 8 emotions defined in [31].

The authors of the dataset presented in [28] considered the top 1.000 positive and negative words in SentiWordNet [29] as keywords to search and crawl over 586.000 images from Flickr. The list of image URLs as well as the collected including title, image resolution, description and the list of the associated tags is available for comparisons [http://www.l3s.de/ minack/flickr-sentiment].

The Geneva Affective Picture Database (GAPED) [74] dataset includes 730 pictures labelled considering negative (e.g. images depicting human rights violation scenes), positive (e.g. human and puppies) as well as neutral pictures which show static objects. All dataset images have been rated considering the valence, arousal, and the coherence of the scene.

The dataset is available for research purposes [http://www.affective-sciences.org/home/research/materials-and-online-research/research-material/].

In 2013 Borth *et al.* [32] proposed a very large dataset (∼ 0.5 million) of pictures gathered from social media and labelled with ANP concepts. Furthermore, they proposed Twitter benchmark dataset which includes 603 tweets with photos. It is intended for evaluating the performance of automatic sentiment prediction using features of different modalities (text only, image only, and text–image combined). This dataset has been used by most of the state-of-the-art works as evaluation benchmark for Visual Sentiment Analysis, especially when the designed approaches involve the use of ML methods such as in [34, 40], for instance due to the large scale of this dataset.

Marchewka *et al.* [75] presented the Nencki Affective Picture System (NAPS), a dataset of high-quality photos depicting people, faces, animals, objects and landscapes. The photos have been manually rated by 204 participants using a 3D approach and a 9-point rating scale for each dimension.

The Emotion6 dataset, presented and used in [41], has been built considering the Elkman's 6 basic emotion categories [36]. The number of images is balanced over the considered categories and

Table 2 Main benchmark datasets for Visual Sentiment Analysis. Some datasets contains several additional information and annotations. All the listed datasets are publicly available

Year	Dataset	Size	Labelling	Social media	Polarity	Additional metadata
1999	IAPS [70]	716 photos	pleasure, arousal and dominance	✗	✗	✗
2008	Yanulevskaya *et al.* [73]	369 photos	awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
2010	affective image classification dataset [30]	228 paintings 807 photos	awe, amusement, contentment, excitement, disgust, anger, fear, sad	✗	✗	✗
2010	Flickr-sentiment [28]	586.000 Flickr photos	positive, negative	✓	✓	✓
2011	GAPED [74]	730 pictures	positive, negative, neutral	✗	✓	✗
2013	VSO [32]	0.5 M Flickr Photos 603 Twitter images	- adjective-noun pairs - positive or negative	✓	✓	✓
2014	NAPS [75]	1.356 Photos	valence, arousal and approach–avoidance	✗	✗	✗
2015	Emotion6 [41]	1.980 Flickr photos	- valence–arousal score - seven emotions distribution	✓	✗	✗
2015	You *et al.* [40]	1.269 Twitter images	positive, negative	✓	✓	✗
2016	IESN [76]	1 M Flickr photos	- valence, arousal, dominance - awe, amusement, contentment, excitement, disgust, anger, fear, sad - positive, negative	✓	✓	✓
2016	CrossSentiment [46]	90.139 Flickr photos 65.439 Instagram images	positive, negative, neutral	✓	✓	✗
2017	T4SA [51]	1.5 M Twitter images	positive, negative, neutral	✓	✓	✗

the emotions associated with each image is expressed as a probability distribution instead of as a single dominant emotion.

In [40] You *et al.* proposed a dataset with 1269 Twitter images labelled into positive or negative by 5 different annotators. Given the subjective nature of sentiment, this dataset has the advantage to be manually labelled by human annotators, differently than other datasets that have been created collecting images by automatic systems based on textual tags or predefined concepts such as the VSO dataset used in [32].

The work in [76] introduced a large scale dataset named Image-Emotion-Social Net (IESN), including over 1 million images crawled from Flickr, labelled considering either the dimensional (i.e. valence, arousal, and dominance) and the categorical (i.e. Mikels' categories [31]) schemes. An important contribute of this work is the inclusion of information related to about 8.000 users and their interactions in the social platform (e.g. preferences, comments, etc.). In [46] the authors crawled two large sets of social pictures from Instagram and Flickr (CrossSentiment). The list of labelled Instagram and Flickr image URLs is available on the Web [http://mm.doshisha.ac.jp/senti/CrossSentiment.html].

Vadicamo *et al.* [51] crawled ~ 3 M tweets from July to December 2016. The collected tweets have been filtered considering only the ones written in English and including at least an image. The sentiment of the text extracted from the tweets has been classified using a polarity classifier based on a paired LSTM-SVM architecture. The data with the most confident prediction have been used to determine the sentiment labels of the images in terms of positive, negative, and neutral. The resulting Twitter for Sentiment Analysis dataset (T4SA) consists of ~ 1 M tweets and related ~ 1.5 M images.

Datasets such as GAPED, NAPS and IAPS rely on emotion induction. These kinds of datasets are very difficult to be built in large scale and maintained over time. The ML techniques and the recent deep learning methods are able to obtain impressive results as long as these systems are trained with very large-scale datasets (e.g. VSO [32]). Such datasets can be easily obtained by exploiting the social network platforms by which people share their pictures every day. These datasets allowed the extensive use of ML systems that requires large-scale datasets. This furthered the building of very large datasets such as T4SA in the last few years. Table 2 summarises the main dataset just reported with details about the number of images, the source (e.g. Social Platform, paintings, etc.) and the labelling options.

3.3 Features

One of the most difficult steps for the design of a Visual Sentiment Analysis system, and in general for the design of a data analysis approach is the selection of the data features that better encode the information that the system is aimed to infer. Image features for Visual Sentiment Analysis can be categorised within three levels of semantics:

- *Low-level features* – These features describe distinct visual phenomena in an image mainly related in some way to the colour values of the image pixels. They usually include generic features such as colour histograms, HOG, GIST. In the context of Visual Sentiment Analysis, previous works can be exploited to extract particular low-level features derived from proper studies on art and perception theory. These studies suggested that some low-level features, such as colours and texture can be used to express the emotional effect of an image [30].
- *Mid-level features* – This group of features bring more semantic, thus they are more interpretable and have stronger associations with emotions [77]. One example is given by the scene-based 102D feature defined in [34]. Furthermore, many of the aforementioned works on Visual Sentiment Analysis exploit the 1200D mid-level representation given by the 1200 ANPs classifiers defined by Borth *et al.* [32].
- *High-level features* – These features describe the semantic concepts shown in the images. Such a feature representation can be obtained by using pre-trained classification methods or semantic embeddings [46].

In 2010 Machajdik and Hanbury [30] performed an intensive study on image emotion classification by properly combining the use of several low and high visual features. These features have been obtained by exploiting concepts from and art theory [12, 16], or exploited in image retrieval [78] and image classification [9, 17] tasks. They selected 17 visual features, categorised into four groups:

- *Colour:* mean saturation and brightness, 3D emotion representation by Valdez and Mehrabian [16], hue statistics, colourfulness measure according to [17], number of pixels of each of the 11 basic colours [79], Itten contrast [12], colour histogram designed by Wei-ning *et al.* [9].
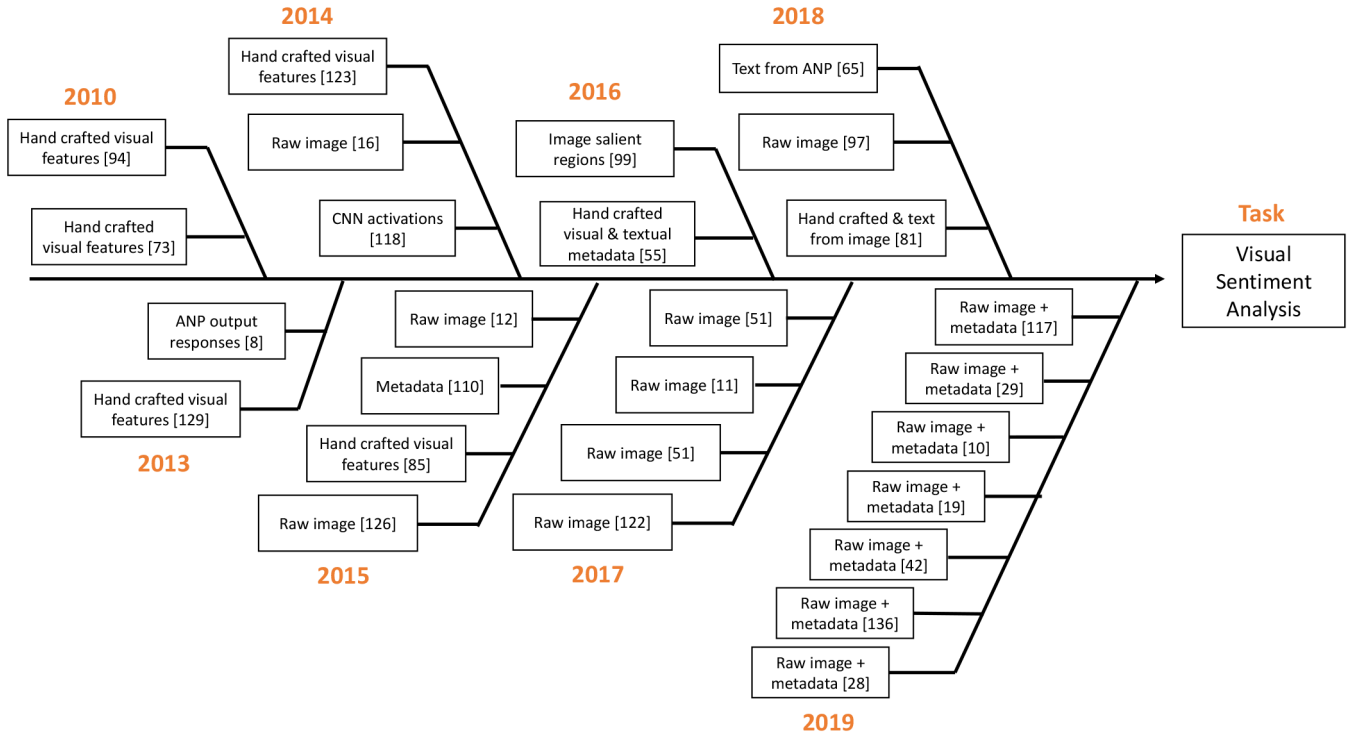
Fig. 5 *Temporal evolution of the inputs considered by the state-of-the-art works in Visual Sentiment Analysis*

- *Texture:* wavelet textures for each HSB channel, features by Tamura *et al.* [80], and features based on GLCM (i.e. correlation, contrast, homogeneity, and energy for the HSB channels).
- *Composition:* the number of resulting segments obtained after the application of a waterfall segmentation (denoted as 'level of detail' in [30]), depth-of-field (DOF) [17], statistics on the line slopes by using the Hough transform (denoted as 'dynamics'), rule of thirds.
- *Content:* number of detected front faces, number of the biggest face pixels, count of skin pixels, ratio of the skin pixels over the face size.

Most of the mentioned works in Visual Sentiment Analysis combine huge number of hand-crafted visual features. Although all the exploited features have been proven to have a direct influence on the perceived emotion by previous studies, there is not agreement about which of them give the most of the contribution on the aimed task. Besides the selection of proper hand-crafted features, designed with the aim to encode the sentiment content conveyed by images, there are other kinds of approaches that lean on representation learning techniques based on deep learning [37, 39, 40]. By employing such representation methods image features are learn from the data. This avoids the designing of a proper feature for the task of Visual Sentiment Analysis, because the system automatically learns how to extract the needed information from the input data. These methods require a huge amounts of labelled training data, and an intensive learning phase, but obtain better performances in general.

Another approach, borrowed from the image retrieval methods, consists on combining textual and visual information through multimodal embedding systems [46, 57, 61, 62]. In this case, features taken from different modalities (e.g. visual, textual, etc.) are combined to create a common vector space in which the correlations between projections of the different modalities are maximised (i.e. an embedding space).

In the diagram depicted in Fig. 5, the related works described in Section 2 are arranged temporally. For each work, the diagram reports the input features considered by the proposed system. It is possible to notice that early works were mostly based on hand-crafted visual features until 2016. Then, most of the works implemented systems able to automatically learn meaningful features from the input raw image (see branch 2017 in Fig. 5).

Finally, the mentioned works published in 2019 combine the contribution of both the raw image and the associated metadata, as they can exploit large dataset built considering images crawled from social media platforms, which offer several additional textual metadata associated to images.

So far, there is not an established strategy to select the visual features that allows to address the problem. Most of the previous exploited features demonstrated to be useful, but recent results on Visual Sentiment Analysis suggest that it's worth investigating the use of representation learning approaches such as CNNs and multimodal embeddings that are able to embody the contribute of multiple sources of information.

4 Problem analysis

In this section, we propose a formulation of the problem, which highlights the related issues and the key tasks of Visual Sentiment Analysis. This allows to better focus the related sub-issues which form the Visual Sentiment Analysis problem and support the designing of more robust approaches. Moreover, to address the overall structure of the problem is useful to suggest a common framework helping researchers to design more robust approaches. Starting from the definition of the Sentiment Analysis problem applied to the natural language text given by Liu [81], we propose to generalise the definition in the context of Visual Sentiment Analysis.

Text-based Sentiment Analysis can be performed considering different levels of detail:

- At the *document level* the task is to classify whether a whole document (i.e. the whole input) expresses a positive or negative sentiment. This model works on the underlying assumption that the whole input discusses only one topic.
- At the *sentence level* the task is to find each phrase within the input document and determine if each sentence expresses a positive or a negative (or neutral) sentiment.
- The *entity and aspect level* performs fine-grained analysis by considering all the opinions expressed in the input document and defining a sentiment score (positive or negative) for each detected target.

Similarly, if the subject of the analysis is an image, we can

- Consider a Sentiment Analysis evaluation for the whole image. These systems work with global image features (e.g. colour histograms, saturation, brightness, colourfulness, colour harmony, etc.).
- Consider an image as a composition of several sub-images according to its specific content. A number of sub-images are extracted and the sentiment analysis is performed on each sub-image obtained by exploiting methods such as multi-object detection, image segmentation, objectness extraction [82].
- Define a set of image aspects, in terms of low level features, each one associated with a sentiment polarity based on previous studies [30, 73]. This is essentially the most fine-grained analysis to be considered.

When a system aims to perform Sentiment Analysis on some textual content, basically it is looking for the opinions in the content and extracting the associated sentiment. An opinion consists of two main components: a target (or topic), and a sentiment. The opinions can be taken from more than one person, this means that the system has to take into account also the opinion holder. Furthermore, opinions can change over time, thus also the time an opinion is expressed has to be taken into account. According to Liu [81], an opinion (or sentiment) is a quintuple

$$(e_i, a_{ij}, s_{ijkh}, h_k, t_l) \tag{1}$$

where e_i is the name of an entity, a_{ij} is an aspect related to e_i, s_{ijkh} is the sentiment score with respect to the aspect a_{ij}, h_k is the opinion holder, and t_l is the time when the opinion is expressed by the opinion holder h_k. The sentiment score s_{ijkh} can be expressed in terms of polarity, considering positive, negative or neutral polarity; or with different levels of intensity. The special aspect 'GENERAL' is used when the sentiment is expressed for the whole entity. In this case, either the entity e_i and the aspect a_{ij} represent the opinion target.

This definition is given in the context of opinion analysis applied on textual contents which express positive or negative sentiments. In the case of Sentiment Analysis applied on visual contents, there are some differences. Indeed, when the input is a text, Sentiment Analysis can easily lean on context and semantic information extracted directly from the text. Thus the problem is to be considered in the NLP domain. When the input is an image, because of the *affective gap* between visual content representations and semantic concepts such as human sentiments, the task to associate the visual features with sentiment labels or polarity scores results challenging. Such *affective gap* can be defined as:

'the lack of coincidence between the measurable signal properties, commonly referred to as features, and the expected affective state in which the user is brought by perceiving the signal' [83].

In the following paragraphs each of the sentiment components previously defined (entity, aspect, holder, and time) are discussed in the context of Visual Sentiment Analysis.

4.1 Entity and aspects

The entity is the subject (or target) of the analysis. In the case of Visual Sentiment Analysis, the entity is the input image. In general, an entity can be viewed as a set of 'parts' and 'attributes'. The set of the entity's parts, its attributes, plus the special aspect 'GENERAL' forms the set of the aspects. This structure can be transferred to the visual domain considering different levels of visual features. Indeed, as mentioned above, in the case of visual contents, Sentiment Analysis can be performed considering different level of visual detail. The most general approach performs Sentiment Analysis considering the whole image, this corresponds to apply a Visual Sentiment Analysis method on the 'GENERAL' aspect. The parts of an image can be defined by considering a set of sub-images. This set can be obtained by exploiting several Computer Vision techniques, such as background/foreground extraction, image segmentation, multi-object recognition or dense captioning [84, 85]. The attributes of an image regards its aesthetic
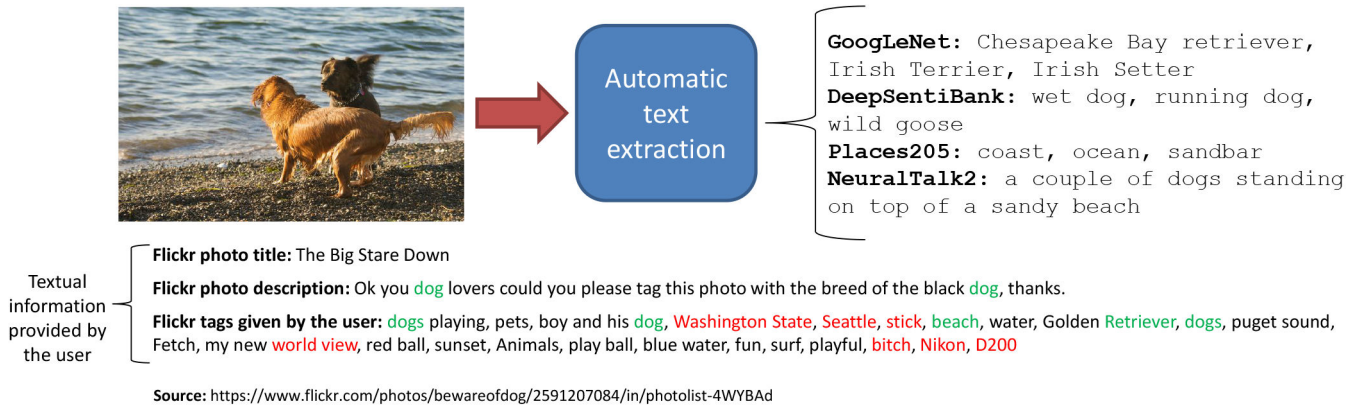
quality features, often obtained by extracting low-level features. Exploiting this structured image hierarchy, a sentiment score can be achieved for each aspect. Finally, the scores are combined to obtain the sentiment classification (e.g. data can be used as input features of a regression model).

Instead of representing the image parts as a set of sub-images, an alternative approach can rely on a textual description of the depicted scene. The description of a photo can be focused on a specific task of image understanding. By changing the task, we can obtain different descriptions of the same image from different points of view. Then, these complementary concepts can be combined to obtain the above described structure. Most of the existing works in sentiment analysis of social media exploit textual information manually associated with images by performing a textual Sentiment Analysis. Although the text associated with social images is widely exploited in the state-of-the-art to improve the semantics inferred from images, it can be a very noisy source because it is provided by the users; the reliability of such input is often based on the capability and the intent of the users to provide textual data that are coherent with respect to the visual content of the image. There is no guarantee that the subjective text accompanying an image is useful. In addition, the tags associated with social images are often selected by users with the purpose to maximise the retrieval and/or the visibility of such images by the platform search engine. In Flickr, for instance a good selection of tags helps to augment the number of views of an image, hence its popularity in the social platform. These information are hence not always useful for sentiment analysis. Starting from this observation, the authors of [57] presented a work on Image Polarity Prediction exploiting Objective Text extracted directly from images, and experimentally compared such text with respect to the Subjective (i.e. user provided) text information commonly used in the state-of-the-art approaches. For a deeper analysis, a comprehensive treatise of image tag assignment is presented in [86].

As discussed in [87], the semantic of an image can be expressed by means of an object category (i.e. a class). However, the tags provided by users usually include several additional terms, related to the object class, coming from a larger vocabulary. As an alternative, the semantic could be expressed by using multiple keywords corresponding to scenes, object categories, or attributes.

Fig. 6 shows an example taken from [57]. The textual information below the image is the text provided by the Flickr's user. Namely the photo title, the description and the tags are usually the text that can be exploited to make inferences on the image. This example shows how the text can be very noisy with respect to any task aimed to understand the sentiment that can be evoked by the picture. Indeed the title is used to describe the tension between the depicted dogs, whereas the photo description is used to ask a question to the community. Furthermore, most of the provided tags include misleading text such as geographical information (i.e. Washington State, Seattle), information related to the camera (i.e. Nikon, D200), objects that are not present in the picture (i.e. boy, red ball, stick) or personal considerations of the user (i.e. my new world view). Moreover, in the subjective text there are many redundant terms (e.g. dog). Another drawback of the text associated with social images is that two users can provide rather different information about the same picture, either in quality and in quantity. Finally, there is not guarantee that such text is present; this is an intrinsic limit of all Visual Sentiment Analysis approaches exploiting subjective text.

Starting from the aforementioned observations about the user provided text associated with social images, the paper in [57] exploits an objective aspect of the textual source that comes directly from the understanding of the visual content. This text is achieved by employing a set of deep learning models trained to accomplish different visual inference tasks on the input image. At the top right part of Fig. 6, the text automatically extracted with different scene understanding methods is shown. In this case, the inferred text is very descriptive and each model provides distinctive information related to objects, scene, context, etc. The objective text extracted by the three different scene understanding methods has a pre-defined structure, therefore all the images have

GoogLeNet: Chesapeake Bay retriever, Irish Terrier, Irish Setter
DeepSentiBank: wet dog, running dog, wild goose
Places205: coast, ocean, sandbar
NeuralTalk2: a couple of dogs standing on top of a sandy beach

Textual information provided by the user

Flickr photo title: The Big Stare Down

Flickr photo description: Ok you dog lovers could you please tag this photo with the breed of the black dog, thanks.

Flickr tags given by the user: dogs playing, pets, boy and his dog, Washington State, Seattle, stick, beach, water, Golden Retriever, dogs, puget sound, Fetch, my new world view, red ball, sunset, Animals, play ball, blue water, fun, surf, playful, bitch, Nikon, D200

Source: https://www.flickr.com/photos/bewareofdog/2591207084/in/photolist-4WYBAd

Fig. 6 *The paper in [57] extracts the text describing a given image by exploiting four different deep learning architectures. The considered architectures are used to extract text related to objects, scene, and image description. The figure shows also the text associated to the image by the user (i.e. title, description, and tags) in the bottom. The subjective text presents very noisy words which are highlighted in red. The words that appear in both sources of text are highlighted in green*

Fig. 7 *Picture of the World Trade Center, taken in 1990*

the same quantity of textual objective information. Such approach provides an alternative user-independent source of text which describes the semantic of images, useful to address the issues related to the inherent subjectivity of the text associated to images.

Several papers faced the issues related to the subjective text associated with images, such as tag refinement and completion [86, 88–90], which aims at alleviating the number of noisy tags and enhancing the number of informative tags by modelling the relationship between visual content and tags.

4.2 Holder

Emotions are subjective, they are affected by several factors such as gender, individual background, age, environments, etc. However, emotions also have the property of stability [91]. This means that the average emotion response of a statistically large set of observers is stable and reproducible. The stability of emotion response enables researchers to generalise their results, when obtained on large datasets.

Almost all the works in Visual Sentiment Analysis ignore the sentiment holder, or implicitly consider only the sentiment of the image publisher. In this context, at least two holders can be defined: the image owner and the image viewer. Considering the example of an advertising campaign, where the owner is the advertising company and the viewers are the potential customers, it's crucial the study and analysis the connection between the sentiment intended by the owner and the actual sentiment induced to the viewers.

These days, the social media platforms provide a very powerful mean to retrieve real-time and large-scale information of people reactions toward topics, events and advertising campaigns. The work in [38] is the first that distinguishes publisher affect (i.e. intent) and viewer affect (i.e. reaction) related to the visual content. This branch of research can be useful to understand the relation between the affect concepts of the image owner and the evoked viewer ones, allowing new user centric applications.

Zhao *et al.* [76] proposed a method to predict the personalised emotion induced by an image for each individual viewer (i.e. user

in the social platform). To this aim, the authors considered several contextual factors such as the social context, temporal evolution and the location. This work required the definition of an *ad-hoc* dataset which includes information about a large scale of involved users, beside the set of images. User profiling helps personalisation, which is very important in the field of recommendation systems. The insights that could be obtained from such a research branch can be useful for several business fields, such as advertisement and User Interface (UI) design. And the community of user interface designers started to take into account the emotional effect of the user interfaces toward users who are interacting with a website, product, or brand. The work in [92] discusses about methods to measure user's emotion during an interface interaction experience, with the aim to assess the interface design emotional effect. Progresses in this field promote the definition of new design approaches such as Emotional UI [93], aimed to exploit the emotions conveyed by visual contents. Indeed, emotions have been traditionally considered to be something that the design evoked, now they represent something that drives the design process. While so far, designers focused on 'user friendly' design (i.e. interfaces easy to use), now they need to focus on design that stimulates and connects the product with users deeply.

Although the interesting cues discussed in this paragraph, currently the development of Visual Sentiment Analysis algorithms that concern the sentiment holder find difficulties due the lack of specific datasets. In this context, the huge data shared on the social media platforms can be exploited to better understand the relationships between the sentiment of the two main holders (i.e. owner/publisher and viewer/user) through their interactions.

4.3 Time

Although almost all the aforementioned works ignore this aspect, the emotion evoked by an image can change depending on the time. This sentiment component can be ignored the most of times, but in specific cases is determinant. Moreover, is very difficult to collect a dataset related to the changes in the emotion evoked by images over time. For example, the sentiment evoked by an image depicting the World Trade Center (Fig. 7) is presumably different if the image is shown before or after 9/11.

Although there are not working on Visual Sentiment Analysis that analyse the changes of image sentiments over time, due to the specificity of the task and the lack of image datasets, there are several works that exploit the analysis of images over time focused on specific cognitive and psychology applications. As an example, the work in [94] employed a statistical framework to detect depression by analysing the sequence of photos posted on Instagram. The findings of this paper suggest the idea that variations in individual psychology reflect in the use social media by the users, hence they can be computationally detected by the analysis of the user's posting history.

In [95] the authors studied which objects and regions of an image are positively or negatively correlated with memorability,

allowing to create memorability maps for each image. This work provides a method to estimate the memorability of images from many different classes. To collect human memory scores, the adopted experimental procedure consists of showing several occurrences of the same images at variable time intervals. The employed image dataset has been created by sampling images from a number of existing datasets, including images evoking emotions [30].

5 Challenges

So far we discussed about the current state-of-the-art in Visual Sentient Analysis, describing the related issues, as well as the different employed approaches and features. This section aims to introduce some additional challenges and techniques that can be investigated.

5.1 Popularity

One of the most common application fields of Visual Sentiment Analysis is related to social marketing campaigns. In the context of social media communication, several companies are interested to analyse the level of people engagement with respect to social posts related to their products. This can be measured as the number of post's views, likes, shares or by the analysis of the comments. These information can be further combined with web search engine and companies website visits statistics, to find correlations between social advertising campaigns and their aimed outcomes (e.g. brand reputation, website/store visits, product dissemination and sale, etc.) [96–98].

The popularity of an image is a difficult quantity to define, hence to measure or infer. However, human beings are able to predict what visual contents other people will like in specific contexts (e.g. marketing campaigns, professional photography). This suggests that there are some common appealing factors into images. So far, researchers have been trying to gain insights into what features make an image popular.

As mentioned in the previous sections, the quality of the text associated with images is often pre-processed in order to avoid noisy text. On the other hand, the users who want to increase the reach of their published contents are used to associate popular tags to their images, regardless their relevance to the image content. This is motivated by the fact that image associated tags are used by the image search engines in these platforms, so the use of popular tags highs the reach of the pictures. Therefore, in this context, the tags associated to images become a crucial factor to understand the popularity of images in social platforms. For instance, Yamasaki *et al.* [99] proposed an algorithm to estimate the social popularity of images uploaded on Flickr by using only text tags. Other features should be also taken into account such as number of user followers and groups, which represent the reach capability of the user. These factors make the task of popularity prediction very different from the task of sentiment polarity classification in the selection of features, methods and measures of evaluation.

The work in [26] considers the effect of 16 features in the prediction of the image popularity. Specifically, they considered image context (i.e. day, time, season and acquisition settings), image content (i.e. image content provided by detectors of scenes, faces and dominant colours), user context and text features (i.e. image tags). The authors cast the problem as a binary classification by splitting the dataset between images with high and low popularity measures. As popularity measures the authors considered the views and the comments counts. Their study highlights that comments are more predictable than views, hence comments are more correlated with the studied features. The experimental results show that the accuracy values achieved only considering textual features (i.e. tags) outperform the performances of the classification based on other features and the combinations of them, for both comments and views counts classifications.

In 2014 Khosla *et al.* [25] proposed a log-normalised popularity score that has been then commonly used in the community. Let c_i be a measure of the engagement achieved by a social media item (e.g. number of likes, number of views, number of shares, etc.), also known as popularity measure. The popularity score of the *i*th item is defined as follows:

$$score_i = \log\left(\frac{c_i}{T_i} + 1\right) \tag{2}$$

where T_i is the number of days since the uploading of the image on the social platform. Equation (2) normalises the number of interactions reached by an image by dividing the engagement measure by the time. However, the measures c_i related to social posts are cumulative values as they continuously collect the interactions between users and the social posts during their time on-line. Therefore, this normalisation will penalise social media contents published in the past with respect to more recent contents, especially when the difference between the dates of posting is high. Indeed, the most of the engagement obtained by a social media item is achieved in the first period, then the engagement measures become more stable. There are very few works which take into account the evolution of the image popularity over time. For example, the study presented in [100] shows that photos have obtained most of their engagement within the first 7 days since the date of upload. However, this study is focused on Flickr, and each social platform has its own mechanisms to show contents to users.

In [25] the authors analysed the importance of several images and social cues that lead to high or low values of popularity. In particular, they considered the relevance of user context features (e.g. mean views, number of photos, number of contacts, number of groups, average groups' members, etc.), image context features (e.g. title length, description length, number of tags), as well as image features (e.g. GIST, LBP, BoW colour patches, CNN activations features, etc.). Is interesting to notice that, differently than several works on sentiment polarity prediction in which the text concerning the images (i.e. title, description and tags included in the post) is semantically analysed in order to achieve sentiment related insights on the image content, in this work only the length of the text associated to the images is considered.

Cappallo *et al.* [101] addressed the popularity prediction problem as a ranking task by exploiting a latent-SVM objective function defined such that the ranking of the popularity scores between pairs of images is maintained. They considered the number of views and comments for Flickr images and the number of re-tweets and favourites for Twitter.

The problem of image popularity prediction is also addressed in [24], whose experiments suggest that some sentiment ANPs defined in VSO [32] have a correlation with popularity.

In [102] the authors considered the number of likes achieved within the first hour after the image posting (early popularity) to predict the popularity after a day, a week or a month. This study has been performed on Instagram images and the dataset is publicly available [http://www1bpt.bridgeport.edu/~jelee/sna/pred.html]. The images are categorised as popular or not popular considering a popularity threshold obtained with the Pareto principle (80–20%). Three features representing information that is retrieved within the first hour of image upload (i.e. early information) are evaluated: social context (based on the user's number of followers), image semantics (based on image caption and NLP), and early popularity in the first hour. The binary classification is performed by using a Gaussian Naive Bayes Model. The experimental results show that the early popularity feature significantly outperforms the other evaluated features. Furthermore, the authors compared the proposed semantic feature with the features proposed by McParlane *et al.* [26] for the task of popularity binary classification considering the MIR-1M Flickr dataset [103], obtaining better accuracy rates.

Most of the works addressing the problem of popularity prediction follow a very similar pipeline. First, a set of interesting features that have been demonstrating correlation with the images sentiment or popularity is selected. Then a model for each distinctive feature is trained to understand the predictive capability of each feature. In the above discussed works, the popularity prediction task is cast as a ranking or a regression problem. Therefore, the exploited algorithms are ranking SVM and latent SVM in the case of ranking, and SVR for regression. Then, the

features are combined with the aim to improve the performances of the method. The evaluation is usually quantified through the Spearman's correlation coefficient.

Although the task of image popularity prediction is rather new, there are interesting datasets available for the development of massive learning systems (i.e. deep neural networks). The Micro-Blog Images 1 Million (MBI-1M) dataset is a collection of 1M images from Twitter, along with accompanying tweets and metadata. The dataset was introduced by the work in [101]. A subset of the Trec 2013 micro-blog track tweets collection [104] has been selected.

The MIR-1M dataset [103] is a collection of 1M photos from Flickr. These images have been selected considering the interestingness score used by Flickr to rank images.

The Social Media Prediction (SMP) dataset is a large-scale collection of social posts, recently collected for the ACM Multimedia 2017 SMP Challenge [Challenge webpage: https://social-media-prediction.github.io/MM17PredictionChallenge]. This dataset consists of over 850 K posts and 80 K users, including photos from VSO [32] as well as photos collected from personal users' albums [105–107]. In particular, the authors aimed to record the dynamic variance of social media data. Indeed, the social media posts in the dataset are obtained with temporal information (i.e. posts sequentiality) to preserve the continuity of post sequences. Two challenges have been proposed:

- *Popularity prediction:* the task is to predict a popularity measure defined for the specific social platform (e.g. number of photo's views on Flickr) of a given image posted by a specific user.
- *Tomorrow's top prediction:* given a set of photos and the data related to the past photo sharing history, the task is to predict the top-*n* popular posts (i.e. ranking problem over a set of social posts) on the social media platform in the next day.

The SMP dataset includes features such as unique picture id (pid) and associated user id (uid). From these information one can extract almost all the user and photo related data available in Flickr. Some metadata of the picture and user-centred information are also included in the dataset. Moreover, the popularity scores (as defined in (2)) are provided. The SMP dataset furthered the development of time aware popularity prediction methods, which exploit time information to define new image representation spaces used to infer the image popularity score at a precise time or at pre-defined time scales. Li *et al.* [108] extracted multiple time-scale features from a set of timestamps related to the photo post. As instance, the timestamp 'postdate' is used to define several features with different time scales: 'season of year', 'month of year', etc. The framework presented in [109] exploits an ensemble learning method to combine the outputs of an SVR and a CART (Classification And Regression Tree) models, previously trained to estimate the popularity score. The models have been trained by exploiting features extracted from user's information, image meta-data and visual aesthetic features extracted from the image. In particular, the authors take into account the post duration (i.e. the number of days the image was posted), the upload time, day, and month.

5.1.1 Popularity dynamics: Equation (2) normalises the number of interactions reached by an image by dividing the engagement measure by the time. However, the measures c_i related to social posts are cumulative values as they continuously collect the interactions between users and the social posts during their time on-line. Therefore, this normalisation will penalise social media contents published in the past with respect to more recent contents, especially when the difference between the dates of posting is high. In other words, such a score does not take into account the evolution of the popularity over time. As a consequence, two images with the same popularity dynamic could be ranked differently, depending on the download time. In particular, the most of the engagement obtained by a social media item is achieved in the first period, then the engagement measures become more stable. For example, the study presented in [100] shows that photos have obtained most of their engagement within the first 7 days

since the date of upload. However, this study is focused on Flickr, and each social platform has its own mechanisms to show contents to users. There are very few works which take into account the evolution of the image popularity over time (i.e. the dynamics of the image popularity). The authors of [110] presented a work on image popularity dynamic prediction. An important contribution is the definition of a dataset of $\sim 20\,K$ Flickr images, crawled within the first 2 h from posting, and their daily engagement scores for 30 days. The approach in [110] considers two different properties of a sequence: the sequence shape describes the trend of the popularity, whereas the sequence scale gives a degree of popularity in the whole period (i.e. the maximum value of engagement achieved after 30 days). The authors performed a detailed analysis of the task, demonstrating that for a given sequence of popularity, the scale and shape properties are unrelated. Such a study defined a set of 50 typical shapes for the 30-days sequences that can be used as prototypes for the popularity dynamics, as well as a method to predict the shape and the scale of an image, given a set of post features. At test time, the two outputs are combined to infer the whole popularity sequence. Since the employed features are available at the moment of the post, this method is able to predict the whole sequence of popularity over 30 days with a daily granularity even before the image is posted.

5.2 Image virality

A recent emerging task, closely related to image popularity, is the prediction of the level of virality of an image. The image virality is defined as the quality of a visual content (i.e. images or videos) to be rapidly and widely spread on social networks [111]. Differently than popularity, the virality score takes into account also the number of resubmission of an image by different users. Therefore, images that became popular when they are posted, but not reposted, are not considered to be viral [112]. These often involve images which content itself is less relevant, but are related to current events that drawn attention to the image in a specific period such as a flash news, or a tragedy. The work in [111] focused on understanding the influence of image parts on its virality. In particular, the authors presented a method for the task of simultaneously detection and localicasion of virality in images. The detection consists on the prediction of the virality score on an image. The localisation aims to detect which areas in an image are responsable for making the image viral, this allows to produce an heatmap which highlights the relevant areas of the input image.

5.3 Relative attributes

As discussed in previous sections, several Visual Sentiment Analysis works aim to associate an image one sentiment label over a set of emotional categories or attributes. However, given a set of images that have been assigned to the same emotional category (e.g. joy), it would be interesting to determine their ranking with respect the specific attribute. Such a technique could suggest, for example if a given image *A* conveys more 'joy' than another image *B*. For this purpose, several works on relative attributes can be exploited [113–116]. Furthermore, a ground truth dataset can be built by exploiting human annotators. Given a pair of images, the annotator is requested to indicate which image is closer to the attribute. In this way it's possible to obtain a proper ranking for each sentiment attribute.

5.4 Common sense

With the aim to reduce the affective and the cognitive gap between images and sentiments conveyed by them, we further need to encode the 'affective common-sense'. A Halloween picture can be classified as a negative image of an automatic system which considers the image semantics, however the knowledge of the context (i.e. Halloween) should affect the semantic concepts conveyed by the picture, hence its interpretation. This corresponds to the 'common-sense knowledge problem' in the field of knowledge representation, which is a sub-field of Artificial Intelligence. Clearly, besides inferential capabilities, such an intelligent programme needs a representation of the knowledge. By

	Image [twemoji]		Unicode codepoint		Occurrences [5...max]		Position [0...1]		Neg [0...1]		Neut [0...1]		Pos [0...1]		Sentiment score [-1...+1]		Sentiment bar (c.i. 95%)	Unicode name
	😂		0x1f602		14622		0.805		0.247		0.285		0.468		0.221			FACE WITH TEARS OF JOY
	😍		0x1f60d		6359		0.765		0.052		0.219		0.729		0.678			SMILING FACE WITH HEART-SHAPED EYES
	😭		0x1f62d		5526		0.803		0.436		0.220		0.343		-0.093			LOUDLY CRYING FACE

Fig. 8 *Some examples of the Emoji Sentiment Ranking scores and statistics obtained by the study conducted in [124]. The sentiment bar (tenth column) shows the proportion of negativity, neutrality, and positivity of the associated emoji*

observing that is very difficult to build a Sentiment Analysis system that may be used in any context with accurate classification prediction, Agrawal *et al.* [117] considered contextual information to determine the sentiment of text. Indeed, in this paper is proposed a model based on common-sense knowledge extracted from ConceptNet [118] ontology and context information. Although this work addresses the problem of Sentiment Analysis applied on textual data, as discussed above, the knowledge of the context related to what an image is depicting should affect its interpretation. Moreover, such results on textual analysis can be transferred to the visual counterpart. Furthermore, emerging approaches based on the Attention mechanism could be exploited to add such a context. The Attention mechanism is a recent trend in deep learning, it can be viewed as a method for making the artificial neural network work better by letting the network know where to look as it is performing its task. For example, in the task of image captioning, the attention mechanism tells the network roughly which pixels to pay attention to when generating the text [119, 120].

5.5 Emoticon/Emoji

In this section, we discuss about the possibility to exploit text ideograms, such emoticons and emoji, in the task of Sentiment Analysis on both visual and textual contents. An emoticon is a textual shorthand that represents a facial expression. The emoticons have been introduced to allow the writer to express feelings and emotions with respect to a textual message. It helps to express the correct intent of a text sentence, improving the understanding of the message. The emoticons are used to emulate visual cues in textual communications with the aim to express or explicitly clarify the writer's sentiment. Indeed, in real conversations the sentiment can be inferred from visual cues such as facial expressions, pose and gestures. However, in textual based conversations, the visual cues are not present.

The authors of [121] tried to understand if emoticons could be useful as well on the textual Sentiment Analysis task. In particular, they investigated the role that emoticons play in conveying sentiment and how they can be exploited in the field of Sentiment Analysis. The authors manually labelled 574 emoticons as positive or negative, and combined this emoticon–lexicon with the text based Sentiment Analysis to perform document polarity classification considering both sentence and paragraph levels. A step further the emoticon, is represented by the emoji. An emoji is an ideogram representing concepts such as weather, celebration, food, animals, emotions, feelings, and activities, besides a large set of facial expressions. They have been developed with the aim to allow more expressive messages. Emojis have become extremely popular in social media platforms and instant messaging systems. For example, in March 2015, Instagram reported that almost half of the texts on its platform contain emojis [122].

In [123], the authors exploited the expressiveness carried by emoji, to develop a system able to generate an image content description in terms of a set of emoji. The focus of this system is to use an emoji as a means for image retrieval and exploration. Indeed, it allows to perform an image search by means of an emoji-based query. This approach exploits the expressiveness conveyed by emoji, by leaning on the textual description of these ideograms (see the tenth column in Fig. 8). The work in [125] studied the ways in which an emoji can be related to other common modalities such as text and images, in the context of multimedia research.

This work also presents a new dataset that contains examples of both text–emoji and image–emoji relationships. Most of them contains also strong sentiment properties.

In [124] the authors presented a sentiment emoji lexicon named Emoji Sentiment Ranking. In this paper, the sentiment properties of the emojis have been deeply analysed, and some interesting conclusions have been highlighted. For each emoji, the *Emoji Sentiment Ranking* provides its associated positive, negative, and neutral scores, represented by values between −1 and +1. The authors also proposed a visual tool, named sentiment bar, to better visualise the sentiment properties associated to each emoji (see Fig. 8). The data considered in this analysis consists of 1.6 million labelled tweets. This collection includes text written in 13 different languages. The authors found that the sentiment scores and ranking associated to emojis remain stable among different languages. This property is very useful to overcome the difficulties addressed in multi-lingual contexts. This lexicon represents a precious resource for many useful applications [The Emoji Sentiment Ranking scores computed by [124] can be visualised at the following URL: http://kt.ijs.si/data/Emoji_sentiment_ranking/].

The work in [126] presented a dataset of 4M Twitter images with their associated emojis. Based on such large-scale dataset, the authors defined a deep neural model for image–emoji embedding. The proposed system has been evaluated on the tasks of emoji prediction, sentiment analysis and emotion prediction.

The results and the insights obtained in [123–126] and could be combined to exploit the sentiment conveyed by the emoji on the task of Visual Sentiment Analysis, rather than exploiting textual information as in the most of the previous works. For instance, an image could be represented by considering the distribution of the associate emojis as a sentiment feature, taking a cue from the approach presented in [41].

By a few years, Facebook has released a new 'reactions' feature, which allows users to interact with a Facebook post by using one of six emotional reactions (Like, Love, Haha, Wow, Sad, and Angry), instead of just having the option of 'liking' a post. These reactions correspond to a meaningful subset of emoji.

6 Discussion

The papers presented in Section 2 represent the most relevant works in the field fo Visual Sentiment Analysis aimed to infer the emotions evoked by pictures. As previously mentioned, there is not agreement on these works about meaningful input features, benchmark datasets nor emotion categories. For this reason, we further investigated these aspects in details in Section 4. The basic polarity classification task considers two or three polarity levels (i.e. negative, neutral, and positive). However, usually more sophisticated psychology-based emotional models are used, such as the Plutchnik's wheel of emotions [33], the Ekman's 5 basic emotions [36], the 35 impression words [68], or the Mikels' 8 basic emotions [31]. Moreover, the works that aim to infer a distribution over a set of sentiment categories [41, 48, 49] employ variations of Mikels' or Ekman's models. Other works, employ specific categorisations such as the 5-scale sentiment categorisation used in [39], or the ANP annotation introduced in [32].

The employed datasets can be organised into two main categories: psychological based or social media based. Datasets based on psychological studies [30, 70, 73], such as IAPS and GAPED, have different pros and cons. Indeed, the photos have been designed to select stimuli eliciting specific ranges of

emotions, and the data can lean on reliable labelling (i.e. human annotations supported by tools/scientists). However, this kind of dataset requires high efforts to be built and maintained over time. As a consequence, is not possible to build large scale datasets. Moreover, considering the selection process of the photos, they results very different kind of images with respect to the ones shared through social media platforms.

More recent efforts have been spent in building large-scale datasets of images shared through social media platforms (e.g. Flickr, Twitter, Facebook, Instagram, etc.). These media provide an easy way to retrieve images (i.e. automatic crawling from social media platforms) at very large scale. Moreover, these images are often associated with accompanying metadata that provides additional context and help for the automatic labelling of the images (e.g. by exploiting NLP on metadata). However, such automatic labelling is often unreliable (e.g. metadata includes sarcam, personal considerations, misleading text, etc.). Therefore, such large-scale datasets need further filtering processes before their exploitation in Visual Sentiment Analysis systems.

Feature selection is crucial for any data analysis process. Data features are supposed to encode the information that the system will be able to infer. There are three general approaches in the field of Image Sentiment Analysis: find correlations between visual features (e.g. colours, textures, etc.) and sentiment, learn image features from data (e.g. training a CNN), exploiting a multi-modal representation learning approach by combining visual and textual information through multi-modal embedding systems (i.e. defining a new vector space).

7 Summary and future directions

In this paper, we have discussed the main issues and techniques related to Visual Sentiment Analysis. The current state-of-the-art has been analysed in detail, highlighting pros and cons of each approach and dataset. Although this task has been studied for years, the field is still in its infancy. Visual Sentiment Analysis is a challenging task due to a number of factors that have been discussed in this paper.

The results discussed in this study, such as [30], agree that the semantic content has a great impact on the emotional influence of a picture. Images having similar colour histograms and textures could have completely different emotional impacts. As a result, a representation of images which express both the appearance of the whole image and the intrinsic semantic of the viewed scene is needed. Earlier methods in the literature about Visual Sentiment Analysis tried to fill the so-called *affective gap* by designing visual representations. Some approaches build systems trained with human labelled datasets and try to predict the polarity of the images. Other approaches compute the polarity of the text associated to the images (e.g. post message, tags and comments) by exploiting common Sentiment Analysis systems that works on textual contents [29, 127], and try to learn ML systems able to infer that polarity from the associated visual content. These techniques have achieved interesting improvements in the tasks of image content recognition, automatic annotation and image retrieval [87, 128–133]. However, it is impossible to know if such user provided text is related to the image content or to the sentiment it conveys. Moreover, the text associated with images is often noisy. Therefore, the exploitation of such text for the definition of either polarity ground truth or as an input source for a sentiment classifier have to address with not reliable text sources. A specific study on this issue is reported in [57]. Furthermore, some approaches exploit a combination of feature modalities (often called views) to build feature space embeddings in which the correlation of the multi-modal features associated to the images that have the same polarity is maximised [46, 57, 61, 62]. The results achieved by several discussed works suggest that exploiting multiple modalities is mandatory, since the sentiment evoked by a picture is affected by a combination of factors, beside the visual information. This has been further confirmed by recent works that add the attention mechanism to the multi-modality approach [59, 60]. Studies in psychology and art theory suggested some visual features associated with emotions evoked by images. However, the most promising choice is given by representations automatically learned through neural networks, autoencoder and feature embedding methods. These approaches are able to find new feature spaces which capture contributes from the different input factor which the sentiment is affected by. The recent results in representation learning confirm this statement.

To this end, one important contribution is given by the availability of large and robust datasets. Indeed, in this study, we highlighted some issues related to the existing datasets. Modern social media platforms allow the collection of huge amount of pictures with several correlated information. These can be exploited to define either input features and 'ground truth'. However, as highlighted before, these textual information need to be properly filtered and processed, in order to avoid the association of noisy information to the images.

Systems with broader ambitions could be developed to address the new challenges (e.g. relative attributes, popularity prediction, common-sense, etc.) or to focus on new emerging tasks (e.g. image popularity prediction, sentiment over time, sentiment by exploiting ideograms, etc.). For instance, ideograms help people to reduce the gap between real and virtual communications. Thanks to the diffusion of social media platforms, the use of emojis has been growing for years, and they are now integrated into the way people communicate in the digital world. They are commonly used to express user reactions with respect to messages, pictures, or news. Thus, the analysis of such new communication media could help to improve the current state of the art performances.

This paper aimed to give a comprehensive overview of the Visual Sentiment Analysis problem, the relative issues, and the algorithms proposed in the state-of-the-art. Relevant points with practical applications in business fields which would benefit from studies in Sentiment Analysis on visual contents have been also discussed. This survey addresses all the aspects related to Visual Sentiment Analysis, and for each of these, relevant references, datasets and results are presented. It also proposes a critical perspective for each considered issue. The paper aims to serve as a starting point for researchers that want to tackle tasks in the field of Visual Sentiment Analysis and related challenges, as it is presented as a structured and critique review of previous works, description of the available datasets, features and techniques. It also aims to clear the way for new solutions, by suggesting new and interesting techniques and sources of information that can be explored to tackle the problem.

Although the psychological basis (e.g. emotional models) and some techniques are often applied on both images and videos, the presented study is mostly devoted to the analysis of still images. Future extensions of this paper could be devoted to the review of the state-of-the-art in Visual Sentiment Analysis applied on videos. In the case of videos, the temporal dimension and the presence of the audio or other type of temporal signals (e.g. EEG [134]) pave the way for a number of sentiment related tasks on visual contents such as multi-modal continuous emotional state recognition [134] or scene popularity [135, 136], among others. Several datasets designed for these tasks are already available [134–136].

8 References

[1] Liu, B., Zhang, L.: 'A survey of opinion mining and sentiment analysis', in Aggarwal, C.C., Zhai, ChengXiang (Eds.): *Mining text data*, (Springer, Boston, MA, USA, 2012), pp. 415–463

[2] Tumasjan, A., Sprenger, T.O., Sandner, P.G., *et al.*: 'Predicting elections with twitter: What 140 characters reveal about political sentiment', Fourth Int. AAAI Conf. on Weblogs and Social Media, Washington, DC, USA, 2010

[3] Pang, B., Lee, L.: 'Opinion mining and sentiment analysis', *Found. trends Inf. Retr.*, 2008, **2**, (1–2), pp. 1–135

[4] Dohaiha, H.H., Prasad, P., Maag, A., *et al.*: 'Deep learning for aspect-based sentiment analysis a comparative review', *Expert Syst. Appl.*, 2018, **118**, pp. 272–299

[5] Zhang, L., Wang, S., Liu, B.: 'Deep learning for sentiment analysis: a survey', *Wiley Interdiscip. Rev.: Data Min. Knowl. Discov.*, 2018, **8**, (4), p. e1253

[6] Ortis, A., Farinella, G.M., Battiato, S.: 'An overview on image sentiment analysis: methods, datasets and current challenges', Proc. of the 16th Int. Joint Conf. on e-Business and Telecommunications - Volume 1: SIGMAP, INSTICC, Prague, Czech Republic, 2019, pp. 290–300

[7] Colombo, C., Del Bimbo, A., Pala, P.: 'Semantics in visual information retrieval', *IEEE Multimedia*, 1999, **6**, (3), pp. 38–53

[8] Schmidt, S., Stock, W.G.: 'Collective indexing of emotions in images. a study in emotional information retrieval', *J. Am. Soc. Inf. Sci. Technol.*, 2009, **60**, (5), pp. 863–876

[9] Wei-ning, W., Ying-lin, Y., Sheng-ming, J.: 'Image retrieval by emotional semantics: a study of emotional space and feature extraction', IEEE Int. Conf. on Systems, Man and Cybernetics, Taipei, Taiwan, 2006, vol. 4, pp. 3534–3539

[10] Zhao, S., Yao, H., Yang, Y., *et al.*: 'Affective image retrieval via multi-graph learning', Proc. of the 22nd ACM Int. Conf. on Multimedia, Orlando, FL, 2014, pp. 1025–1028

[11] Bradley, M.M.: 'Emotional memory: a dimensional analysis', in van Goozen, S.H.M., van de Poll, N.E., Sergeant, J.A. (Eds.): *Emotions: Essays on Emotion Theory*, (Erlbaum, Hillsdale, 1994), pp. 97–134

[12] Itten, J.: '*The art of color: the subjective experience and objective rationale of color* (John Wiley & Sons Inc, New York, USA, 1973)

[13] Lang, P.J.: 'The network model of emotion: motivational connections', *Perspect. Anger Emotion: Adv. Soc. Cogn.*, 1993, **6**, pp. 109–133

[14] Osgood, C.E.: 'The nature and measurement of meaning', *Psychol. Bull.*, 1952, **49**, (3), p. 197

[15] Russell, J.A., Mehrabian, A.: 'Evidence for a three-factor theory of emotions', *J. Res. Personality*, 1977, **11**, (3), pp. 273–294

[16] Valdez, P., Mehrabian, A.: 'Effects of color on emotions', *J. Exper. Psychol.: Gen.*, 1994, **123**, (4), p. 394

[17] Datta, R., Joshi, D., Li, J., *et al.*: 'Studying aesthetics in photographic images using a computational approach', European Conf. on Computer Vision, Graz, Austria, 2006, pp. 288–301

[18] Joshi, D., Datta, R., Fedorovskaya, E., *et al.*: 'On aesthetics and emotions in scene images: a computational perspective', in Kveraga, K., Bar, M. (Eds.): *Scene vision: making sense of what we see*, (The MIT Press, Cambridge, MA, USA, 2014), p. 241

[19] Joshi, D., Datta, R., Fedorovskaya, E., *et al.*: 'Aesthetics and emotions in images', *IEEE Signal Process. Mag.*, 2011, **28**, (5), pp. 94–115

[20] Marchesotti, L., Perronnin, F., Larlus, D., *et al.*: 'Assessing the aesthetic quality of photographs using generic image descriptors', 2011 Int. Conf. on Computer Vision, Barcelona, Spain, 2011, pp. 1784–1791

[21] Raví, F., Battiato, S.: 'A novel computational tool for aesthetic scoring of digital photography', Proc. of 6th European Conf. on Colour in Graphics, Imaging, and Vision (SPIE-IS&T), Amsterdam, 2012, pp. 1–5

[22] Isola, P., Xiao, J., Torralba, A., *et al.*: 'What makes an image memorable?', IEEE Conf. on Computer Vision and Pattern Recognition, Colorado Springs, CO, USA, 2011, pp. 145–152

[23] Jia, J., Wu, S., Wang, X., *et al.*: 'Can we understand van gogh's mood?: learning to infer affects from images in social networks', Proc. of the 20th ACM Int. Conf. on Multimedia, Nara, Japan, 2012, pp. 857–860

[24] Gelli, F., Uricchio, T., Bertini, M., *et al.*: 'Image popularity prediction in social media using sentiment and context features', Proc. of the 23rd ACM Int. Conf. on Multimedia, Brisbane Australia, 2015, pp. 907–910

[25] Khosla, A., Das Sarma, A., Hamid, R.: 'What makes an image popular?', Proc. of the 23rd Int. Conf. on World Wide Web, Seoul, Korea, 2014, pp. 867–876

[26] McParlane, P.J., Moshfeghi, Y., Jose, J.M.: 'Nobody comes here anymore, it's too crowded; predicting image popularity on flickr', Proc. of Int. Conf. on Multimedia Retrieval, Glasgow, UK, 2014, p. 385

[27] Totti, L.C., Costa, F.A., Avila, S., *et al.*: 'The impact of visual attributes on online image diffusion', Proc. of the 2014 ACM Conf. on Web Science, Bloomington, IN, USA, 2014, pp. 42–51

[28] Siersdorfer, S., Minack, E., Deng, F., *et al.*: 'Analyzing and predicting sentiment of images on the social web', Proc. of the 18th ACM Int. Conf. on Multimedia, Firenze, Italy, 2010, pp. 715–718

[29] Esuli, A., Sebastiani, F.: 'Sentiwordnet: a publicly available lexical resource for opinion mining', Proc. of LREC, Genoa, Italy, 2006, vol.**6**, pp. 417–422

[30] Machajdik, J., Hanbury, A.: 'Affective image classification using features inspired by psychology and art theory', Proc. of the 18th ACM Int. Conf. on Multimedia, Firenze, Italy, 2010, pp. 83–92

[31] Mikels, J.A., Fredrickson, B.L., Larkin, G.R., *et al.*: 'Emotional category data on images from the international affective picture system', *Beh. Res. Meth.*, 2005, **37**, (4), pp. 626–630

[32] Borth, D., Ji, R., Chen, T., *et al.*: 'Large-scale visual sentiment ontology and detectors using adjective noun pairs', Proc. of the 21st ACM Int. Conf. on Multimedia, Barcelona, Spain, 2013, pp. 223–232

[33] Plutchik, R.: 'A general psychoevolutionary theory of emotion', *Theor. Emotion*, 1980, **1**, pp. 3–31

[34] Yuan, J., Mcdonough, S., You, Q., *et al.*: 'Sentribute: image sentiment analysis from a mid-level perspective', Proc. of the Second Int. Workshop on Issues of Sentiment Discovery and Opinion Mining, Chicago, IL, USA, 2013, p. 10

[35] Yang, Y., Jia, J., Zhang, S., *et al.*: 'How do your friends on social media disclose your emotions?', Proc. of the Twenty-Eighth AAAI Conf. on Artificial Intelligence (AAAI'14), Québec City, Québec, Canada, 2014, pp. 306–312

[36] Ekman, P., Friesen, W.V., O'Sullivan, M., *et al.*: 'Universals and cultural differences in the judgments of facial expressions of emotion', *J. Personality Soc. Psychol.*, 1987, **53**, (4), p. 712

[37] Chen, T., Borth, D., Darrell, T., *et al.*: 'Deepsentibank: Visual sentiment concept classification with deep convolutional neural networks. ArXiv preprint arXiv:1410.8586 2014

[38] Chen, Y.-Y., Chen, T., Hsu, W.H., *et al.*: 'Predicting viewer affective comments based on image content in social media', Proc. of Int. Conf. on Multimedia Retrieval, Glasgow, UK, 2014, p. 233

[39] Xu, C., Cetintas, S., Lee, K.-C., *et al.*: Visual sentiment prediction with deep convolutional neural networks. arXiv preprint arXiv:1411.5731, 2014

[40] You, Q., Luo, J., Jin, H., *et al.*: 'Robust image sentiment analysis using progressively trained and domain transferred deep networks', Proc. of the Twenty-Ninth AAAI Conf. on Artificial Intelligence (AAAI'15), Austin, TX, USA, 2015, pp. 381–388

[41] Peng, K.-C., Chen, T., Sadovnik, A., *et al.*: 'A mixed bag of emotions: model, predict, and transfer emotion distributions', Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 860–868

[42] Wang, Y., Wang, S., Tang, J., *et al.*: 'Unsupervised sentiment analysis for social media images', Proc. of the 24th Int. Conf. on Artificial Intelligence (IJCAI'15), Buenos Aires, Argentina, 2015, pp. 2378–2379

[43] Campos, V., Salvador, A., Giró-i Nieto, X., *et al.*: 'Diving deep into sentiment: understanding fine-tuned cnns for visual sentiment prediction', Proc. of the 1st Int. Workshop on Affect & Sentiment in Multimedia (ASM '15), New York, NY, USA, 2015, pp. 57–62

[44] Jou, B., Chen, T., Pappas, N., *et al.*: 'Visual affect around the world: a large-scale multilingual visual sentiment ontology', Proc. of the 23rd ACM Int. Conf. on Multimedia, Brisbane, Australia, 2015, pp. 159–168

[45] Sun, M., Yang, J., Wang, K., *et al.*: 'Discovering affective regions in deep convolutional neural networks for visual sentiment prediction', IEEE Int. Conf. on Multimedia and Expo., Seattle, WA, USA, 2016, pp. 1–6

[46] Katsurai, M., Satoh, S.: 'Image sentiment analysis using latent correlations among visual, textual, and sentiment views', IEEE Int. Conf. on Acoustics, Speech and Signal Processing, Shanghai, China, 2016, pp. 2837–2841

[47] Yu, F.X., Cao, L., Feris, R.S., *et al.*: 'Designing category-level attributes for discriminative visual recognition', Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Portland, OR, USA, 2013, pp. 771–778

[48] Yang, J., Sun, M., Sun, X.: 'Learning visual sentiment distributions via augmented conditional probability neural network', AAAI, San Francisco, CA, USA, 2017, pp. 224–230

[49] Jufeng Yang, M.S., She, D.: 'Joint image emotion classification and distribution learning via deep convolutional neural network', Proc. of the Twenty-Sixth Int. Joint Conf. on Artificial Intelligence, (IJCAI-17), Melbourne, Australia, 2017, pp. 3266–3272

[50] Campos, V., Jou, B., Nieto, X.G.i.: 'From pixels to sentiment: fine-tuning CNNS for visual sentiment prediction', *Image Vis. Comput.*, 2017, **65**, pp. 15–22, Multimodal Sentiment Analysis and Mining in the Wild Image and Vision Computing

[51] Vadicamo, L., Carrara, F., Cimino, A., *et al.*: 'Cross-media learning for image sentiment analysis in the wild', Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Honolulu, HI, USA, 2017, pp. 308–317

[52] Simonyan, K., Zisserman, A.: Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014

[53] Zhou, B., Lapedriza, A., Xiao, J., *et al.*: 'Learning deep features for scene recognition using places database', Advances in Neural Information Processing Systems, Montreal, Quebec, Canada, 2014, pp. 487–495

[54] Li, Z., Fan, Y., Liu, W., *et al.*: 'Image sentiment prediction based on textual descriptions with adjective noun pairs', *Multimedia Tools Appl.*, 2018, **77**, (1), pp. 1115–1132

[55] Thelwall, M., Buckley, K., Paltoglou, G., *et al.*: 'Sentiment strength detection in short informal text', *J. Assoc. Inf. Sci. Technol.*, 2010, **61**, (12), pp. 2544–2558

[56] Song, K., Yao, T., Ling, Q., *et al.*: 'Boosting image sentiment analysis with visual attention', *Neurocomputing*, 2018, **312**, pp. 218–228

[57] Ortis, A., Farinella, G.M., Torrisi, G., *et al.*: 'Visual sentiment analysis based on objective text description of images', 2018 Int. Conf. on Content-Based Multimedia Indexing (CBMI), La Rochelle, France, 2018, pp. 1–6

[58] Felicetti, A., Martini, M., Paolanti, M., *et al.*: 'Visual and textual sentiment analysis of daily news social media images by deep learning', Int. Conf. on Image Analysis and Processing, Trento, Italy, 2019, pp. 477–487

[59] Zhu, X., Cao, B., Xu, S., *et al.*: 'Joint visual-textual sentiment analysis based on cross-modality attention mechanism', Int. Conf. on Multimedia Modeling, Thessaloniki, Greece, 2019, pp. 264–276

[60] Huang, F., Zhang, X., Zhao, Z., *et al.*: 'Image–text sentiment analysis via deep multimodal attentive fusion', *Knowl.-Based Syst.*, 2019, **167**, pp. 26–37

[61] Campos, V., Giro-i Nieto, X., Jou, B., *et al.*: 'Sentiment concept embedding for visual affect recognition', in Alameda-Pineda, X., Ricci, E., Sebe, N. (Eds.): *Multimodal behavior analysis in the wild*, (Elsevier, Long Beach, CA, USA, 2019), pp. 349–367

[62] Fortin, M.P., Chaib-draa, B.: 'Multimodal sentiment analysis: a multitask learning approach', Proc. of the 8th Int. Conf. on Pattern Recognition Applications and Methods, Prague, Czech Republic, 2019

[63] Corchs, S., Fersini, E., Gasparini, F.: 'Ensemble learning on visual and textual data for social image emotion classification', *Int. J. Machine Learn. Cybern.*, 2019, **10**, (8), pp. 2057–2070

[64] Wu, L., Qi, M., Jian, M., *et al.*: 'Visual sentiment analysis by combining global and local information', *Neural Process. Lett.*, 2019, pp. 1–13

[65] Zhao, S., Ding, G., Huang, Q., *et al.*: 'Affective image content analysis: a comprehensive survey', Int. Joint Conf. on Artificial Intelligence, 2018, pp. 5534–5541

[66] Soleymani, M., Garcia, D., Jou, B., *et al.*: 'A survey of multimodal sentiment analysis', *Image Vis. Comput.*, 2017, **65**, pp. 3–14

[67] Li, Z., Fan, Y., Jiang, B., *et al.*: 'A survey on sentiment analysis and opinion mining for social multimedia', *Multimedia Tools Appl.*, 2019, **78**, (6), pp. 6939–6967

[68] Hayashi, T., Hagiwara, M.: 'Image query by impression words-the iqi system', *IEEE Trans. Consum. Electron.*, 1998, **44**, (2), pp. 347–352

[69] Hanjalic, A., Xu, L.-Q.: 'Affective video content representation and modeling', *IEEE Trans. Multimedia*, 2005, **7**, (1), pp. 143–154

[70] Lang, P.J., Bradley, M.M., Cuthbert, B.N.: 'International affective picture system (IAPS): technical manual and affective ratings', in *The center for research in psychophysiology*, University of Florida, Gainesville, FL, 1999

[71] Shaver, P., Schwartz, J., Kirson, D., *et al.*: 'Emotion knowledge: further exploration of a prototype approach', *J. Personality Soc. Psychol.*, 1987, **52**, (6), p. 1061

[72] Izard, C.E.: 'Basic emotions, relations among emotions, and emotion-cognition relations', 1992

[73] Yanulevskaya, V., Van Gemert, J., Roth, K., *et al.*: 'Emotional valence categorization using holistic image features', 15th IEEE Int. Conf. on Image Processing, San Diego, CA, USA, 2008, pp. 101–104

[74] Dan-Glauser, E.S., Scherer, K.R.: 'The geneva affective picture database (gaped): a new 730-picture database focusing on valence and normative significance', *Beh. Res. Meth.*, 2011, **43**, (2), pp. 468–477

[75] Marchewka, A., Żurawski, Ł., Jednoróg, K., *et al.*: 'The nencki affective picture system (naps): introduction to a novel, standardized, wide-range, high-quality, realistic picture database', *Beh. Res. Meth.*, 2014, **46**, (2), pp. 596–610

[76] Zhao, S., Yao, H., Gao, Y., *et al.*: 'Predicting personalized emotion perceptions of social images', Proc. of the 24th ACM Int. Conf. on Multimedia, Amsterdam, The Netherlands, 2016, pp. 1385–1394

[77] Zhao, S., Gao, Y., Jiang, X., *et al.*: 'Exploring principles-of-art features for image emotion recognition', Proc. of the 22nd ACM Int. Conf. on Multimedia, Orlando, FL, USA, 2014, pp. 47–56

[78] Stottinger, J., Banova, J., Ponitz, T., *et al.*: 'Translating journalists' requirements into features for image search', 15th Int. Conf. on Virtual Systems and Multimedia, Vienna, Austria, 2009, pp. 149–153

[79] Van de, J., eijer, W, Schmid, C., *et al.*: 'Learning color names from real-world images', 2007 IEEE Conf. on Computer Vision and Pattern Recognition, Minneapolis, MN, USA, 2007, pp. 1–8

[80] Tamura, H., Mori, S., Yamawaki, T.: 'Textural features corresponding to visual perception', *IEEE Trans. Syst. Man Cybern.*, 1978, **8**, (6), pp. 460–473

[81] Liu, B.: 'Sentiment analysis and opinion mining', *Synth. Lect. Hum. Lang. Technol.*, 2012, **5**, (1), pp. 1–167

[82] Alexe, B., Deselaers, T., Ferrari, V.: 'Measuring the objectness of image windows', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2012, **34**, (11), pp. 2189–2202

[83] Hanjalic, A.: 'Extracting moods from pictures and sounds: towards truly personalized tv', *IEEE Signal Process. Mag.*, 2006, **23**, (2), pp. 90–100

[84] Karpathy, A., Fei-Fei, L.: 'Deep visual-semantic alignments for generating image descriptions', The IEEE Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, June 2015

[85] Karpathy and, A., Fei-Fei, L.: 'Deep visual-semantic alignments for generating image descriptions', Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Boston, MA, USA, 2015, pp. 3128–3137

[86] Li, X., Uricchio, T., Ballan, L., *et al.*: 'Socializing the semantic gap: a comparative survey on image tag assignment, refinement, and retrieval', *ACM Comput. Surv. (CSUR)*, 2016, **49**, (1), p. 14

[87] Gong, Y., Ke, Q., Isard, M., *et al.*: 'A multi-view embedding space for modeling internet images, tags, and their semantics', *Int. J. Comput. Vis.*, 2014, **106**, (2), pp. 210–233

[88] Sang, J., Xu, C., Liu, J.: 'User-aware image tag refinement via ternary semantic analysis', *IEEE Trans. Multimed.*, 2012, **14**, (3), pp. 883–895

[89] Wu, L., Jin, R., Jain, A.K.: 'Tag completion for image retrieval', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2013, **35**, (3), pp. 716–727

[90] Xu, H., Wang, J., Hua, X.-S., *et al.*: 'Tag refinement by regularized lda', Proc. of the 17th ACM Int. Conf. on Multimedia, New York, USA, 2009, pp. 573–576

[91] Wang, W., He, Q.: 'A survey on emotional semantic image retrieval', 15th IEEE Int. Conf. on Image Processing, San Diego, CA, USA, 2008, pp. 117–120

[92] Lockner, D., Bonnardel, N., Bouchard, C., *et al.*: 'Emotion and interface design', Proc. of the 2014 Ergonomie et Informatique Avancée Conf.-Design, Ergonomie et IHM: 'Quelle Articulation Pour la Co-Conception de l'interaction, New York, USA, 2014, pp. 33–40

[93] Kazim, S.: An introduction to emotive UI, April 2016, accessed 17 April 2018

[94] Reece, A.G., Danforth, C.M.: 'Instagram photos reveal predictive markers of depression', *EPJ Data Sci.*, 2017, **6**, (1), p. 15

[95] Khosla, A., Xiao, J., Torralba, A., *et al.*: 'Memorability of image regions', *Adv. Neural Inf. Process. Syst.*, 2012, pp. 305–313

[96] Goeldi, A.: Website network and advertisement analysis using analytic measurement of online social media content, 5 July 2011. US Patent 7,974,983

[97] Hanna, R., Rohm, A., Crittenden, V.L.: 'We're all connected: the power of the social media ecosystem', *Bus. Horiz.*, 2011, **54**, (3), pp. 265–273

[98] Trusov, M., Bucklin, R.E., Pauwels, K.: 'Effects of word-of-mouth versus traditional marketing: findings from an internet social networking site', *J. Mark.*, 2009, **73**, (5), pp. 90–102

[99] Yamasaki, T., Sano, S., Aizawa, K.: 'Social popularity score: 'predicting numbers of views, comments, and favorites of social photos using only annotations', Proc. of the First Int. Workshop on Internet-Scale Multimedia Management, Orlando, FL, USA, 2014, pp. 3–8

[100] Valafar, M., Rejaie, R., Willinger, W.: 'Beyond friendship graphs: a study of user interactions in flickr', Proc. of the 2nd ACM Workshop on Online Social Networks, Barcelona, Spain, 2009, pp. 25–30

[101] Cappallo, S., Mensink, T., Snoek, C.G.: 'Latent factors of visual popularity prediction', Proc. of the 5th ACM on Int. Conf. on Multimedia Retrieval, Shanghai, People's Republic of China, 2015, pp. 195–202

[102] Almgren, K., Lee, J., Kim, M., *et al.*: 'Predicting the future popularity of images on social networks', Proc. of the The 3rd Multidisciplinary Int. Social Networks Conf. on SocialInformatics 2016, Data Science 2016, Union, NJ, USA, 2016, p. 15

[103] Huiskes, M.J., Thomee, B., Lew, M.S.: 'New trends and ideas in visual concept detection: 'the mir flickr retrieval evaluation initiative', Proc. of the Int. Conf. on Multimedia Information Retrieval, Vancouver, British Columbia, Canada, 2010, pp. 527–536

[104] Lin, J., Efron, M.: 'Overview of the trec2013 microblog track'. Technical report, 2013

[105] Wu, B., Cheng, W.-H., Zhang, Y., *et al.*: 'Time matters: multi-scale temporalization of social media popularity', Proc. of the 2016 ACM on Multimedia Conf. (ACM MM), Amsterdam, The Netherlands, 2016

[106] Wu, B., Cheng, W.-H., Zhang, Y., *et al.*: 'Sequential prediction of social media popularity with deep temporal context networks', Int. Joint Conf. on Artificial Intelligence (IJCAI), Melbourne, Australia, 2017

[107] Wu, B., Mei, T., Cheng, W.-H., *et al.*: 'Unfolding temporal dynamics: predicting social media popularity using multi-scale temporal decomposition', Proc. of the Thirtieth AAAI Conf. on Artificial Intelligence (AAAI), Phoenix, AZ, USA, 2016

[108] Li, L., Situ, R., Gao, J., *et al.*: 'A hybrid model combining convolutional neural network with xgboost for predicting social media popularity', Proc. of the 2017 ACM on Multimedia Conf. (MM '17), New York, NY, USA, 2017, pp. 1912–1917

[109] Hidayati, S.C., Chen, Y.-L., Yang, C.-L., *et al.*: 'Popularity meter: an influence-and aesthetics-aware social media popularity predictor', Proc. of the 2017 ACM on Multimedia Conf., Mountain View, CA, USA, 2017, pp. 1918–1923

[110] Ortis, A., Farinella, G.M., Battiato, S.: 'Predicting social image popularity dynamics at time zero', *IEEE Access*, 2019, **7**, pp. 1–1

[111] Alameda-Pineda, X., Pilzer, A., Xu, D., *et al.*: 'Viraliency: 'pooling local virality', 2017 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Honolulu, HI, USA, July 2017, pp. 484–492

[112] Deza, A., Parikh, D.: 'Understanding image virality', 2015 IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), Boston, MA, USA, June 2015, pp. 1818–1826

[113] Altwaijry and, H., Belongie, S.: 'Relative ranking of facial attractiveness', IEEE Workshop on Applications of Computer Vision (WACV), Washington, DC, USA, 2013, pp. 117–124

[114] Fan, Q., Gabbur, P., Pankanti, S.: 'Relative attributes for large-scale abandoned object detection', Proc. of the IEEE Int. Conf. on Computer Vision, Sydney, Australia, 2013, pp. 2736–2743

[115] Parikh, D., Grauman, K.: 'Relative attributes', IEEE Int. Conf. on Computer Vision, Barcelona, Spain, 2011, pp. 503–510

[116] Yu, A., Grauman, K.: 'Just noticeable differences in visual attributes', Proc. of the IEEE Int. Conf. on Computer Vision, Washington, DC, USA, 2015, pp. 2416–2424

[117] Agarwal, B., Mittal, N., Bansal, P., *et al.*: 'Sentiment analysis using common-sense and context information', *Comput. Intell. Neurosci.*, 2015, **2015**, p. 9

[118] Liu, H., Singh, P.: 'Conceptnet — a practical commonsense reasoning tool-kit', *BT Technol. J.*, 2004, **22**, (4), pp. 211–226

[119] Xu, K., Ba, J., Kiros, R., *et al.*: 'Show, attend and tell: neural image caption generation with visual attention', Int. Conf. on Machine Learning, Lille, France, 2015, pp. 2048–2057

[120] You, Q., Jin, H., Wang, Z., *et al.*: 'Image captioning with semantic attention', Proc. of the IEEE Conf. on Computer Vision and Pattern Recognition, Las Vegas, NV, USA, 2016, pp. 4651–4659

[121] Hogenboom, A., Bal, D., Frasincar, F., *et al.*: 'Exploiting emoticons in sentiment analysis', Proc. of the 28th Annual ACM Symp. on Applied Computing, Las Vegas, NV, USA, 2013, pp. 703–710

[122] Dimson, T.: Emojineering part 1: Machine learning for emoji trends. https://instagram-engineering.com/emojineering-part-1-machine-learning-for-emoji-trendsmachine-learning-for-emoji-trends-7f5f9cb979ad, 2015, accessed 17 April 2018

[123] Cappallo, S., Mensink, T., Snoek, C.G.: 'Image2emoji: zero-shot emoji prediction for visual media', Proc. of the 23rd ACM Int. Conf. on Multimedia, Brisbane Australia, 2015, pp. 1311–1314

[124] Kralj Novak, P., Smailović, J., Sluban, B., *et al.*: 'Sentiment of emojis', *PLOS ONE*, 2015, **10**, (12), pp. 1–22

[125] Cappallo, S., Svetlichnaya, S., Garrigues, P., *et al.*: 'The new modality: emoji challenges in prediction, anticipation, and retrieval', *IEEE Trans. Multimed.*, 2018, **21**, (2), pp. 402–415. Pending minor revision

[126] Al-Halah, Z., Aitken, A., Shi, W., *et al.*: Smile, be happy:) emoji embedding for visual sentiment analysis. arXiv preprint arXiv:1907.06160, 2019

[127] Wilson, T., Wiebe, J., Hoffmann, P.: 'Recognizing contextual polarity in phrase-level sentiment analysis', Proc. of the Conf. on Human Language Technology and Empirical Methods in Natural Language Processing, 2005, pp. 347–354

[128] Fu, Y., Hospedales, T.M., Xiang, T., *et al.*: 'Transductive multi-view embedding for zero-shot recognition and annotation', European Conf. on Computer Vision, Zurich, Switzerland, 2014, pp. 584–599

[129] Gong, Y., Wang, L., Hodosh, M., *et al.*: 'Improving image-sentence embeddings using large weakly annotated photo collections', European Conf. on Computer Vision, Zurich, Switzerland, 2014, pp. 529–545

[130] Guillaumin, M., Verbeek, J., Schmid, C.: 'Multimodal semi-supervised learning for image classification', IEEE Conf. on Computer Vision & Pattern Recognition, San Francisco, CA, USA, 2010, pp. 902–909

[131] Katsurai, M., Ogawa, T., Haseyama, M.: 'A cross-modal approach for extracting semantic relationships between concepts using tagged images', *IEEE Trans. Multimed.*, 2014, **16**, (4), pp. 1059–1074

[132] Li, Z., Liu, J., Tang, J., *et al.*: 'Robust structured subspace learning for data representation', *IEEE Trans. Pattern Anal. Mach. Intell.*, 2015, **37**, (10), pp. 2085–2098

[133] Rasiwasia, N., Costa Pereira, J., Coviello, E., *et al.*: 'A new approach to cross-modal multimedia retrieval', Proc. of the 18th ACM Int. Conf. on Multimedia, Firenze, Italy, 2010, pp. 251–260

[134] Liu, Y.-J., Yu, M., Zhao, G., *et al.*: 'Real-time movie-induced discrete emotion recognition from eeg signals', *IEEE Trans. Affective Comput.*, 2017, **9**, (4), pp. 550–562

[135] Battiato, S., Farinella, G.M., Milotta, F.L., *et al.*: 'Organizing videos streams for clustering and estimation of popular scenes', Int. Conf. on Image Analysis and Processing, Catania, Italy, 2017, pp. 51–61

[136] Ortis, A., Farinella, G.M., D'amico, V., *et al.*: 'Recfusion: automatic video curation driven by visual content popularity', Proc. of the 23rd ACM Int. Conf. on Multimedia, Brisbane Australia, 2015, pp. 1179–1182