International Journal of Multimedia Information Retrieval (2021) 10:111–125
https://doi.org/10.1007/s13735-021-00207-4

REGULAR PAPER

Multimodal news analytics using measures of cross-modal entity and context consistency

Eric Müller-Budack[1] · Jonas Theiner[2] · Sebastian Diering[2] · Maximilian Idahl[2] · Sherzod Hakimov[1] · Ralph Ewerth[1,2]

Received: 31 January 2021 / Accepted: 20 February 2021 / Published online: 28 April 2021
© The Author(s) 2021

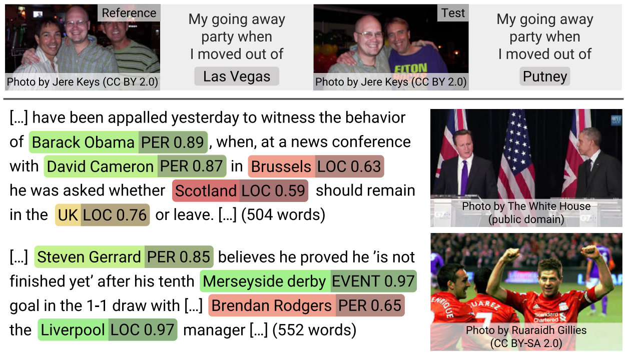
Abstract

The World Wide Web has become a popular source to gather information and news. Multimodal information, e.g., supplement text with photographs, is typically used to convey the news more effectively or to attract attention. The photographs can be decorative, depict additional details, but might also contain misleading information. The quantification of the cross-modal consistency of entity representations can assist human assessors' evaluation of the overall multimodal message. In some cases such measures might give hints to detect fake news, which is an increasingly important topic in today's society. In this paper, we present a multimodal approach to quantify the entity coherence between image and text in *real-world* news. Named entity linking is applied to extract persons, locations, and events from news texts. Several measures are suggested to calculate the cross-modal similarity of the entities in text and photograph by exploiting state-of-the-art computer vision approaches. In contrast to previous work, our system automatically acquires example data from the Web and is applicable to real-world news. Moreover, an approach that quantifies contextual image-text relations is introduced. The feasibility is demonstrated on two datasets that cover different languages, topics, and domains.

Keywords Cross-modal consistency · News analytics · Image-text relations · Image repurposing detection

1 Introduction

With the widespread use and availability of digital environments, the World Wide Web plays an essential role in disseminating information and news. In particular, social media platforms such as *Twitter* allow users to follow worldwide events and news and become a popular source of information [6,35,39]. These news articles often leverage different modalities, e.g., texts and images, to convey information more effectively (Fig. 1). Every modality conveys its specific information, and the combination of modalities enables the communication of a coherent multimodal message. In this regard, photograph content can range from decorative (with little or no information about the news event) over depicting rich information enhancements (important or additional details) to even misleading visual information.


Reference
Photo by Jere Keys (CC BY 2.0)
Test
Photo by Jere Keys (CC BY 2.0)
Photo by The White House (public domain)
Photo by Ruaraidh Gillies (CC BY-SA 2.0)

My going away party when I moved out of Las Vegas

My going away party when I moved out of Putney

[...] have been appalled yesterday to witness the behavior of Barack Obama PER 0.89, when, at a news conference with David Cameron PER 0.87 in Brussels LOC 0.63 he was asked whether Scotland LOC 0.59 should remain in the UK LOC 0.76 or leave. [...] (504 words)

[...] Steven Gerrard PER 0.85 believes he proved he 'is not finished yet' after his tenth Merseyside derby EVENT 0.97 goal in the 1-1 draw with [...] Brendan Rodgers PER 0.65 the Liverpool LOC 0.97 manager [...] (552 words)

Fig. 1 Top: Test and reference images of the *MEIR* dataset [36] and corresponding texts with untampered and tampered entities. Bottom: Two real-world news from *BreakingNews* [33] and outputs of our system (LOCation, PERson, EVENT). The examples show that real-world news have much longer text and refer to many entities. Images are replaced with similar ones due to license restrictions. Original images and full text are linked on the *GitHub* page: https://github.com/TIBHannover/cross-modal_entity_consistency (Color figure online)

✉ Eric Müller-Budack
eric.mueller@tib.eu

✉ Ralph Ewerth
ralph.ewerth@tib.eu

[1] TIB – Leibniz Information Centre for Science and Technology, Hannover, Germany

[2] L3S Research Center, Leibniz University Hannover, Hannover, Germany

According to Bateman [4], the consideration of multimodal relationships such as the semantic coherence and mutual concepts is crucial to understand and evaluate the overall message and meaning. With the rapidly growing amount of news available on the Web, it is becoming an increasingly important task to develop automated systems for information extraction in multimedia content in order to, e.g., evaluate the overall message, facilitate semantic search, or analyze the content with regard to credibility. Measures of cross-modal consistency might also support human assessors and expert-oriented fact-checking efforts such as *PolitiFact*[1] and *Snopes*[2] to identify misinformation or *fake news*.

While part of previous work [16,25,32,44,46] aims at finding measures to model semantic cross-modal relations in order to bridge the semantic gap, approaches on image repurposing detection [21,22,36] check the consistency of named entities mentioned in the text, as illustrated in Fig. 1. Our approach is similar to the task of image repurposing detection since it focuses on the evaluation of cross-modal entity occurrences between image and text. Related approaches [21,22,36] rely on multimodal deep learning techniques that require appropriate datasets of non-manipulated image-text pairs. However, these datasets are hard to collect as they need to be verified for valid cross-modal relations. Besides, the training or reference data provide the source of world knowledge and limit these methods to entities, e.g., persons or locations, that appear in these datasets. Experimental evaluations have been performed on images with short image captions [21,36] or existing metadata [22], which do not reflect real-world characteristics as illustrated in Fig. 1.

In this paper, we present an automatic system that quantifies the cross-modal consistency of entity relations. In contrast to previous work, the system is completely unsupervised and does not rely on any pre-defined reference or training data. To the best of our knowledge, we present a first baseline that is *applicable to real-world news articles* by tackling several news-specific challenges such as the excessive length of news documents, entity diversity, and misleading reference images. The workflow of our system pipeline is as follows: First, we automatically crawl reference images for entities extracted from the text by *named entity linking*. Then, these images serve as input for the visual verification of the entities to the associated news image. In this respect, appropriate computer vision approaches serve as generalized feature extractors. Unlike the more general model for scene/place classification used in [30], we utilize a novel ontology-driven deep learning approach [31] to generate features for event verification. Finally, novel measures for different entity types (*persons*, *locations*, *events*) as well as for a more general news context are introduced to quantify the cross-modal similarity of image and text.

The applications are manifold, ranging from a retrieval system for news with low or high cross-modal semantic correlations to an exploration tool that reveals the relations between image and text as shown in Fig. 1. The feasibility of our approach is demonstrated on a novel large-scale dataset for cross-modal consistency verification that is derived from *BreakingNews* [33]. It contains real-world news articles in *English* and covers different topics and domains. In addition, we have collected articles from *German* news sites to verify the performance in another language. In contrast to previous work, the entities are manipulated with more sophisticated strategies to obtain challenging datasets. Web application, source code, and datasets are publicly available[3].

The remainder of this paper is organized as follows. The framework for automatic verification of cross-modal entity relations as well as contextual relations between image and text is described in Sects. 3 and 4. Section 5 introduces two benchmarks datasets and discusses the experimental results of the proposed approach for document verification and collection retrieval. Section 6 summarizes the section and outlines potential areas of future work.

2 Related work

The analysis of multimodal information such as image and text has attracted researchers from linguistics, semiotics, and computational science for many years. Bateman [4] considers multimodal relationships to be crucial for the interpretation of the overall multimodal message. Linguists, semioticians, and communication scientists [3,13,27,28,40] attempted to assign joint placements of image and text to distinct image-text classes in order to define the interrelations using suitable taxonomies. However, only recently few works attempted to build computational models to quantify the cross-modal relations between image and text. Few approaches explore more general semantic correlations [16,25,32,44,46] to bridge the gap [38] between both modalities.

Also related to our approach are systems for image repurposing detection [21,22,36] that intend to reveal inconsistencies between image-text-pairs with respect to entity representations (*persons*, *locations*, *organizations*, etc.), mainly to identify repurposed multimedia contents that might indicate misinformation. In a more general sense these kind of approaches quantify the *Cross-modal Mutual Information* (CMI) [16,17] of named entities. Jaiswal et al. [21] have learned a multimodal representation of reference packages

[1] https://www.politifact.com/.

[2] https://www.snopes.com/.

[3] Web application: https://labs.tib.eu/newsanalytics Source code & datasets: https://github.com/TIBHannover/cross-modal_entity_consistency.

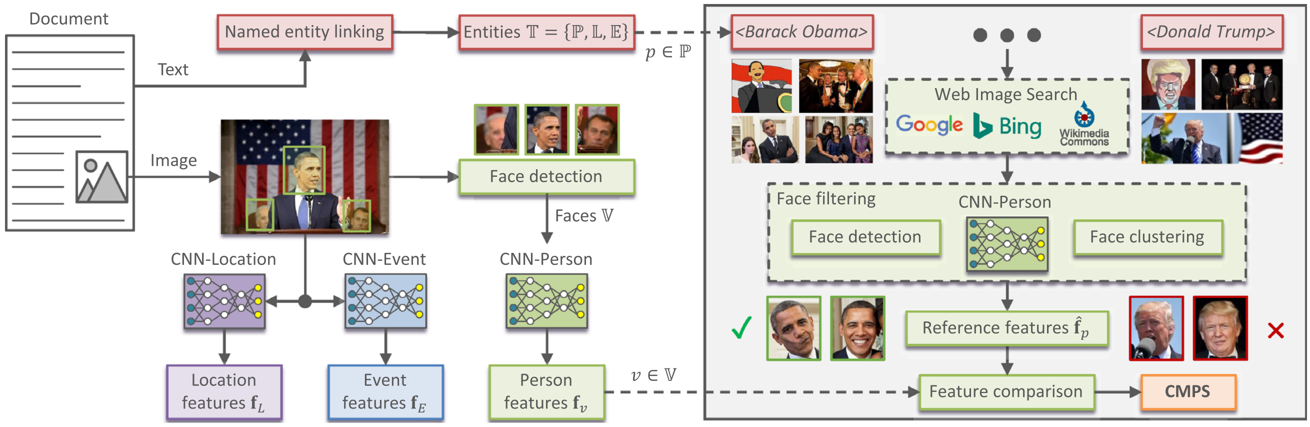


Fig. 2 Workflow of the proposed system to quantify cross-modal entity similarities. Left: Extraction of textual entities $\mathbb{T}$ according to Sect. 3.1, as well as visual features for persons $\mathbb{P}$ (green), locations $\mathbb{L}$ (purple), and events $\mathbb{E}$ (blue) (Sect. 3.2). Right: Workflow to measure the *Cross-modal Person Similarity* (CMPS) between image and text (Sect. 3.3) based on visual evidence crawled from the Web. The same pipeline but without filtering is used for locations and events (Color figure online)

containing an untampered image and a corresponding caption to assess a given document's semantic integrity. Experiments were conducted by replacing one modality, which results in semantically inconsistent image-captions pairs, making them relatively easy to detect. This motivated Sabir et al. [36] to introduce a dataset where specific entities (persons, locations, and organizations) are carefully replaced to generate semantically consistent altered packages. They have also refined the multimodal model using a multitask learning approach that further incorporates geographical information. Jaiswal et al. [22] presented an adversarial neural network that simultaneously trains a bad actor who intentionally counterfeits metadata and a watchdog that verifies multimodal semantic consistency. The system was tested for person verification, location verification, and painter verification of artworks. However, the system is more closely related to approaches for metadata verification [7,8,23,26] as it only verifies the consistency between pairs of images and metadata and does not incorporate any textual information.

Overall, the aforementioned approaches neglect the various challenges of real-world news and applications in terms of the vast amount and variety of entities, incorrect or unrelated reference data as well as outputs of named entity linking tools. They instead rely on pre-defined reference datasets consisting of image-text pairs [21,36] or existing metadata [22] that are (1) closely related (Fig. 1 top), (2) hard to collect automatically, and (3) rather limited and static with respect to the covered entities.

3 Cross-modal entity consistency

In this section, we present a system that automatically verifies the semantic relations in terms of shared entities between pairs of image and text. Verification is realized through measures of cross-modal similarities for different entity types (*persons*, *locations*, and *events*). Based on named entity linking (Sect. 3.1), visual evidence for cross-modal entity occurrences is collected from the Web. Visual features are obtained by appropriate computer vision approaches (Sect. 3.2), which are used in conjunction with measures of cross-modal similarity (Sect. 3.3) to quantify the cross-modal consistency. The workflow is illustrated in Fig. 2.

3.1 Extraction of textual entities

In order to quantify cross-modal relation for specific types of entities, namely *persons*, *locations*, and *events*, named entity recognition and disambiguation is applied to extract a set of named entities $\mathbb{T}$ from the text. We have tried several frameworks such as *AIDA* [18], *NERD* [34], or Kolitsas et al.'s [24] approach. In an initial experiment, we found that combining the output of *spaCy* [19] for named entity recognition and *Wikifier* [5] for named entity linking provide the best results for different languages. Given a named entity recognition system for a specific language, *Wikifier* enables our system to support a large number of 100 languages. We link the entity candidate with the highest *PageRank* according to *Wikifier* for every named entity recognized by *spaCy* to the *Wikidata* knowledge base. Linked entities with a *PageRank* below $1 \cdot e^{-4}$ are neglected due to their low confidence. If *Wikifier* does not provide a linked entity for a given string, the *Wikidata* API function ”*wbsearchentities*” is used for disambiguation.

As shown in Fig. 2, suitable computer vision approaches based on deep learning are applied to extract visual features that are used to quantify the cross-modal entity consistency.

The computer vision model is selected based on the type (*person*, *location*, or *event*) of the named entity. Thus, it is necessary to assign each named entity $t \in \mathbb{T} = \{\mathbb{P}, \mathbb{L}, \mathbb{E}\}$ to one of the entity types to create distinct sets of persons $\mathbb{P}$, locations $\mathbb{L}$, and events $\mathbb{E}$. Although some named entity recognition tools such as *spaCy* [19] automatically predict entity types, they do not make use of the knowledge base information of the linked entities. To handle mistakes of entity type classification by *spaCy* and to discard irrelevant entities such as given names that cannot be linked to a knowledge base, the entity types are re-evaluated using the *Wikidata* information of the linked entities based on the following requirements. For persons only entities that are an instance (*P31*) of *human* (*Q5*) according to *Wikidata* are considered, while for locations a valid *coordinate location* (*P625*) is set as a requirement. This allows us to extract a variety of locations ranging from *continents*, *countries*, and *cities* to specific *landmarks*, *streets*, or *buildings*. For events we instead require an entity to be in a verified list of events[4] according to *EventKG* [10,11]. Entities that do not fulfill any of the aforementioned criteria are neglected. As a result, distinct sets of persons $\mathbb{P}$, locations $\mathbb{L}$, and events $\mathbb{E}$ are extracted from the text that are used to acquire example images from the Web, as explained in Sect. 3.3.

3.2 Extraction of visual features

Our approach is applicable to articles containing multiple images, but we assume that only a single image is present for simplicity. State-of-the-art models are applied to obtain visual image representations.

Person Features: For person verification, we first jointly detect and normalize faces using the *Multi-task Cascaded Convolutional Networks* [45]. An implementation[5] of *Face Net* [37] is used to calculate a feature matrix $\mathbf{F}_{\mathbb{V}}$ that contains the individual feature vectors $\mathbf{f}_v$ of all faces $v \in \mathbb{V}$ found in the image.

Location Features: We employ the *base* (M, f^*) model[6] for geolocalization [29] to obtain a geospatial representation of the article's image. It provides good results across different environmental settings (*indoor*, *natural*, and *urban*). In contrast to the original method, we treat geolocalization as a verification approach and utilize the feature vector $\mathbf{f}_L$ from the penultimate pooling layer of the *ResNet-101* model [14,15].

Event Features: In our initial approach [30], we used a more general image descriptor for scene classification to extract features for events since related approaches for event classification [1,2,43] have not considered many event types that are relevant for news. Recently, we have presented a dataset and ontology-driven deep learning approach for event classification [31]. Unlike previous work, it considers the majority of newsworthy event types such as *natural disasters*, *epidemics*, and *elections*. For this reason, we use this ontology-driven CO_{γ}^{cos} model[7] in the approach described in this paper. The visual event features $\mathbf{f}_E$ are extracted from the last pooling layer of the *ResNet-50* architecture [14,15]. A comparison to the previous approach [30] is conducted in Sect. 5.5.

3.3 Verification of shared cross-modal entities

In this section, we present measures of *Cross-modal Similarity* for different entity types, namely *persons*, *locations*, and *events*. It should be emphasized that we treat each verification task independently. The *Cross-modal Similarity* for different entity types are *not combined* which allows a more detailed and realistic analysis. Referring to Fig. 1 (bottom), please imagine a news article where the image depicts one or several person(s) talking at a conference. While there can be multiple events and locations mentioned in the corresponding text, the news image does not provide any visual cues for their verification. This is typical for news articles since the text usually contains more entities and information. In case of *fake news*, it is common that only one entity type is manipulated to maintain credibility.

3.3.1 Verification of persons

As illustrated in Fig. 2, we first crawl a maximum of k example images using image search engines such as *Google* or *Bing* for each person $p \in \mathbb{P}$ that was extracted from the named entity linking approach presented in Sect. 3.1. Since these images can depict other or several persons, a filtering step is necessary. As described in Sect. 3.2, feature vectors are extracted for each detected face $v \in \mathbb{V}$ in the images. These features are compared with each other using the cosine similarity to perform a hierarchical clustering with a minimal similarity threshold $\tau_{\mathbb{P}}$ as a termination criterion. Consequently, the mean feature vector of the majority cluster is calculated and serves as the reference vector $\hat{\mathbf{f}}_p$ for person p, since it most likely represents the queried person.

Finally, the feature vector $\mathbf{f}_v$ of each face $v \in \mathbb{V}$ detected in the document image is compared to the reference vector $\hat{\mathbf{f}}_p$ of each person $p \in \mathbb{P}$. Several options are available to calculate an overall *Cross-modal Person Similarity* (CMPS)

[4] List of events from *EventKG* [10] available at: http://eventkg.l3s.uni-hannover.de/data/event_list.tsv.

[5] *FaceNet* implementation from *David Sandberg*: https://github.com/davidsandberg/facenet.

[6] *base* (M, f^*) model for geolocation estimation: https://github.com/TIBHannover/GeoEstimation.

[7] CO_{γ}^{cos} model for event classification: https://github.com/TIBHannover/VisE.

such as the mean, $n\%$-quantile, or the max of all comparisons. However, as mentioned above, usually the text contains more entities than the image, and already a single correlation can theoretically ensure credibility. Thus, we define the *Cross-modal Person Similarity* (CMPS) as the maximum similarity among all comparisons according to Eq. (1), since the mean or quantile would require the presence of several or all entities mentioned in the text.

$$\text{CMPS} = \max_{p \in \mathbb{P}, v \in \mathbb{V}} \left(\frac{\mathbf{f}_v \cdot \hat{\mathbf{f}}_p}{||\mathbf{f}_v|| \cdot ||\hat{\mathbf{f}}_p||} \right) \tag{1}$$

3.3.2 Verification of locations and events

In general, we follow the pipeline of person entity verification. The feature vectors of a maximum of k reference images for each location and event mentioned in the text are calculated using the deep learning approach of the respective entity type according to Sect. 3.2. However, while some entities are very specific (e.g., *landmarks*, *sport finals*), others are more general (e.g., *countries*, *international crises*) and can therefore contain diverse example data. This makes a visual filtering based on clustering very complicated since these entities can already contain many visually different subclusters due to high intra-class variations. Thus, the feature vector $\mathbf{f}_L$ (for locations) or $\mathbf{f}_E$ (for events) of the news photograph (Sect. 3.2) is compared to the feature matrix $\hat{\mathbf{F}}_l$ (for locations) or $\hat{\mathbf{F}}_e$ (for events) that contains the features of each reference image crawled for a given location $l \in \mathbb{L}$ or event $e \in \mathbb{E}$ using the cosine similarity according to the following equations:

$$\text{CMLS} = \max_{l \in \mathbb{L}} \Psi \left(\frac{\mathbf{f}_L \cdot \hat{\mathbf{F}}_l}{||\mathbf{f}_L|| \cdot ||\hat{\mathbf{F}}_l||} \right) \tag{2}$$

$$\text{CMES} = \max_{e \in \mathbb{E}} \Psi \left(\frac{\mathbf{f}_E \cdot \hat{\mathbf{F}}_e}{||\mathbf{f}_E|| \cdot ||\hat{\mathbf{F}}_e||} \right) \tag{3}$$

To obtain a *Cross-modal Similarity* value for each entity, an operator function $\Psi : \mathbf{s} \rightarrow [0, 1]$ (e.g., the maximum operator) is applied that reduces the resulting similarity vector $\mathbf{s}$ containing the similarities of all reference image to the news image to a scalar. In the experiments (Sect. 5.3), we evaluate the maximum and several n%-quantiles as potential operator functions. We believe that using a n%-quantile is more robust against incorrect or unrelated entity images in the retrieved reference data. As explained for person verification, we decided to use the maximum *Cross-modal Similarity* among all entities of a given type for both the *Cross-modal Location Similarity* (CMLS) and *Cross-modal Event Similarity* (CMES) of the document.

4 Cross-modal context consistency

In the previous section, we have presented an approach that quantifies the cross-modal consistency for each entity based on reference images crawled from the Web. This approach is not applicable to the quantification of the contextual semantic relation since Web queries are hard to define automatically based on the entire news content. For this reason, we pursued a different direction. We extracted word embeddings from the articles text (Sect. 4.1) as well as the visual probabilities of general scene concepts along with their respective word embeddings (Sect. 4.2) to quantify the *Cross-modal Context Similarity* (CMCS) (Sect. 4.3). An overview is provided in Fig. 3.

4.1 Textual scene context

To retrieve suitable candidates representing the textual (scene) context $\mathbb{C}$, the part-of-speech tagging from *spaCy* [19] is applied to extract all nouns $c \in \mathbb{C}$. They can represent general concepts, such as *politics* or *sports*, as well as *scenes* or *actions*, that might correlate to specific classes, e.g., of a place (scene) classification dataset such as *Places365* [47]. Subsequently, we calculate the word embedding $\mathbf{w}_c$ for each candidate $c \in \mathbb{C}$ using *fastText* [12] as a prerequisite for the cross-modal comparison explained in Sect. 4.3.

4.2 Visual scene context

A *ResNet-50* model[8] [14,15] for scene (place) classification that is trained on 365 places of the *Places365* dataset [47] is applied to predict the visual scene probabilities $\hat{\mathbf{y}}_S$. As for the textual scene context (Sect. 4.1), *fastText* [12] is employed to additionally extract the corresponding word embeddings $\mathbf{w}_s$ of each scene label $s \in \mathbb{S}$. While the scene label such as *beach*, *conference center*, or *church* are rather generic, their word embeddings can be also associated with specific news topics such as *holiday*, *politics*, or *religion*. Both the visual scene probabilities and scene word embeddings are used as visual scene context. The scene labels were manually translated to *German* for the experiments on *German* news articles.

4.3 Cross-modal context similarity

Unlike the cross-modal entity verification, the quantification of the *Cross-modal Context Similarity* (CMCS) does not require any reference images as it is solely based on the textual (Sect. 4.1) and visual scene context (Sect. 4.2) given by the news article. In this regard, we compare the individual

[8] *ResNet-50* model trained with *PyTorch* on *Places365*: https://github.com/CSAILVision/places365.

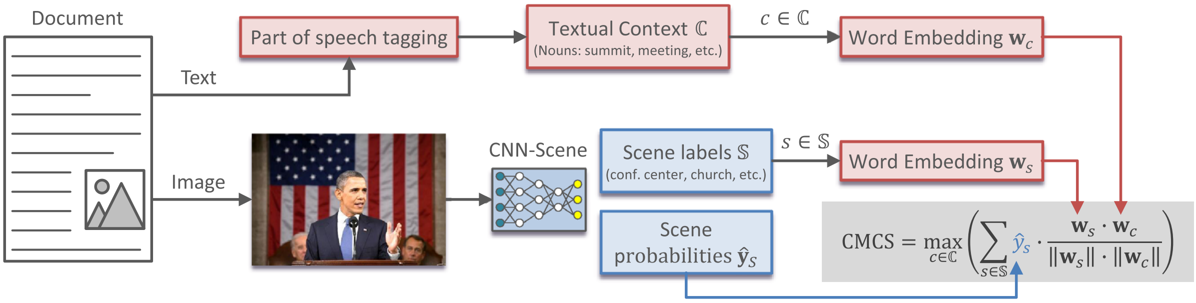

Document
Text
Image
Part of speech tagging
Textual Context $\mathbb{C}$
(Nouns: summit, meeting, etc.)
$c \in \mathbb{C}$
Word Embedding $\mathbf{w}_c$
CNN-Scene
Scene labels $\mathbb{S}$
(conf. center, church, etc.)
$s \in \mathbb{S}$
Word Embedding $\mathbf{w}_s$
Scene probabilities $\hat{\mathbf{y}}_S$
$\mathrm{CMCS} = \max_{c \in \mathbb{C}} \left(\sum_{s \in \mathbb{S}} \hat{y}_s \cdot \frac{\mathbf{w}_s \cdot \mathbf{w}_c}{\|\mathbf{w}_s\| \cdot \|\mathbf{w}_c\|} \right)$

Fig. 3 Workflow of the proposed system to quantify the *Cross-modal Context Similarity* (CMCS). Part-of-speech tagging is applied to extract textual scene context candidates (Sect. 4.1), e.g., nouns from the text. The class names as well as visual probabilities computed by a deep learning approach for place/scene classification define the visual scene context (Sect. 4.3). Finally, the word embeddings from both textual $\mathbf{w}_c$ and visual scene context $\mathbf{w}_s$ are compared and weighted by the visual scene probabilities $\hat{\mathbf{y}}_S = \langle \hat{y}_1, \hat{y}_2, \ldots, \hat{y}_{365} \rangle$ of 365 categories from the *Places365* dataset [47] categories according to Sect. 4.3 to calculate the CMCS (Color figure online)

word embeddings $\mathbf{w}_c$ of each noun $c \in \mathbb{C}$ to the word embeddings $\mathbf{w}_s$ of all 365 scene class labels $s \in \mathbb{S}$ covered by the *Places365* dataset [47] using the cosine similarity. Since only certain scenes are depicted in a news image, these similarities are weighted with the respective visual scene probability $\hat{y}_s$ of a scene class $s \in \mathbb{S}$ to integrate the image information. Finally, the *Cross-modal Context Similarity* (CMCS) is defined as the maximum similarity among all comparisons according to Fig. 3.

5 Experimental setup and results

In this section, we introduce two novel datasets for cross-modal consistency verification (Sect. 5.1). Furthermore, the metrics for evaluation (Sect. 5.2) and parameter selections (Sect. 5.3) are explained in more detail. The performance of the proposed system on real-world news articles is evaluated in Sect. 5.4 and two different deep learning approaches for the quantification of cross-modal event relationships are compared in Sect. 5.5. Finally, the limitations and dependencies of our proposed approach are discussed in Sect. 5.6.

5.1 Datasets

Two real-world news datasets that cover different languages, domains, and topics are utilized for the experiments. They were both manipulated to perform experiments for cross-modal consistency verification. Experiments and comparisons to related work [21,36] on datasets such as *MEIR* [36] are not reasonable since (1) they do not contain public persons or events, and (2) rely on *pre-defined* reference or training data for *given* entities. These restrictions severely limit the application in practice. We propose an automated solution for real-world scenarios that works for public personalities and entities represented in a knowledge base. In the remainder of this section, we introduce the tampering techniques (Sect. 5.1.1) as well as the *TamperedNews* (Sect. 5.1.2) and *News400* (Sect. 5.1.3) datasets, which contain articles written in *English* and *German*, respectively.

Dataset Version 2: Please note that we have noticed a minor problem in the first version of our dataset [30] that affected circa 5% of the linked entities. As a consequence, the results slightly differ from the first version. The repository and datasets have been updated accordingly[3].

5.1.1 Tampering techniques

We have created multiple sets of tampered entities for each document in our datasets. Similar to Sabir et al. [36], we replaced entities extracted from the text at random with another entity of the same type to change semantic relations as little as possible. We also apply more sophisticated tampering techniques as follows. Three additional tampered person sets are created by replacing each untampered person with another person of the same gender (PsG), the same country of citizenship (PsC), or matching both criteria above (PsCG). Locations are replaced by other locations that share at least one parent class (e.g., *country* or *city*) according to *Wikidata* and are located within a *Great Circle Distance* (GCD) of $dmin$ and $dmax$ kilometers ($\mathrm{GCD}_{dmin}^{dmax}$). Three intervals are used to experiment with different spatial resolutions at region-level (GCD_{25}^{200}), country-level (GCD_{200}^{750}), and continent-level ($\mathrm{GCD}_{750}^{2500}$). Similarly, events that share the same parent class (e.g., *sports competition* or *natural disaster*) with the untampered event are used for a second set (EsP) of tampered events. In case no valid candidate for a tampering strategy was available, we have used a random candidate that matched most of the other tampering criteria.

The contextual verification is based on the nouns in the text. Thus, textual tampering techniques are not applicable.

Table 1 Number of test documents $|\mathbb{D}|$, unique entities $\mathbb{T}^*$ in all articles, and mean amount of unique entities $\overline{\mathbb{T}}$ in articles containing a given entity type (for *context* this is the mean amount of nouns as explained in Sect. 4.1) for *TamperedNews* (top) and *News400* (bottom)

Documents	$\|\mathbb{D}\|$	$\mathbb{T}^*$	$\overline{\mathbb{T}}$
TamperedNews dataset			
All (context)	72,561	–	121.40
With person entities	33,695	4772	4.01
With location entities	66,484	3464	4.78
With event entities	15,467	875	1.32
News400 dataset			
All (verified context)	397 (91)	–	137.35
With persons (verified)	320 (116)	413	5.31
With locations (verified)	389 (69)	434	8.83
With events (verified)	166 (31)	39	1.84

Valid image-text relations for *News400* were first manually verified according to Sect. 5.1.3

We instead replaced the image with a random image from all other documents for a first tampered set. We randomly selected similar images (from top-k% with $k \in \{5, 10, 25\}$) to maintain semantic relations to create three more sets. The similarity was computed using feature vectors extracted from a *ResNet* model [14,15] trained on *ImageNet* [9].

5.1.2 TamperedNews dataset

To the best of our knowledge, *BreakingNews* [33] is the largest available corpus with news articles that contain both image and text. It originally covered approximately 100,000 English news articles from 2014 across different domains and a huge variety of topics (e.g., *sports*, *politics*, *healthcare*). We created a subset called *TamperedNews* for cross-modal consistency verification of 72,561 articles for which the news text and image were still available. The entities in these articles were additionally tampered according to Sect. 5.1.1. To discard most irrelevant entities, only persons and locations mentioned at least in ten documents and events that occur in at least three documents are considered. Detailed dataset statistics are reported in Table 1.

5.1.3 News400 dataset

To show the capability of our approach for another language and time period, we have used the *Twitter* API to obtain the web links (URLs) of news articles from three popular *German* news websites (faz.net, haz.de, sueddeutsche.de). The texts and main images of the articles were crawled from the URLs. We have gathered 397 news articles containing four different topics (*politics*, *economy*, *sports*, and *travel*) in the period from August 2018 to January 2019. The smaller size of the dataset allowed us to conduct a manual annotation with three experts to ensure valid relationships between image and text. For each document, the annotators verified the presence of at least one person, location, or event in the image as well as in the text and whether the context was consistent in both modalities. Experiments were conducted exclusively on data with valid relations. Again the tampering techniques presented in Sect. 5.1.1 are applied to create the test sets. Due to its smaller size, every entity is considered regardless of how often it appears in the entire dataset. The resulting statistics are shown in Table 1.

5.2 Evaluation tasks and metrics

The evaluation tasks are motivated by potential real-world applications of our system. We propose to evaluate the system for two tasks: (1) document verification and (2) collection retrieval. The system can also be used as an analytics tool to quickly explore cross-modal relations within a document as illustrated in Fig. 1.

5.2.1 Document verification

Please imagine a set of two or more news articles with similar content and imagery but differences in the mentioned entities that might have been tampered by an author with harmful intents. The idea behind this task is to decide which joint pair of image and entities extracted from the news text provides a higher cross-modal consistency. Thus, a document verification can help users to detect the most or least suitable document. We address this task using the following strategy. For each individual document in the dataset, we compare the cross-modal similarities between the news image and the respective set of untampered entities as well as *one* set of tampered entities (e.g, PsG) according to the strategies proposed in Sect. 5.1.1. This allows us to evaluate the impact of different tampering strategies. We report the *Verification Accuracy* (VA) that quantifies how often the untampered entity set has achieved the higher cross-modal similarity to the document's image. Some qualitative examples are shown in Fig. 5. Please note that the image is tampered for the context evaluation instead and that the nouns in the text are considered as "entities".

5.2.2 Collection retrieval

The system can also be leveraged in news collections to retrieve news articles with high or low cross-modal relations to support human assessors to gather the most credible news or possibly fake news (in extreme cases). We therefore consider all $|\mathbb{D}|$ untampered documents as well as $|\mathbb{D}|$ tampered documents applying *one* tampering strategy. The cross-modal similarities are calculated and used to

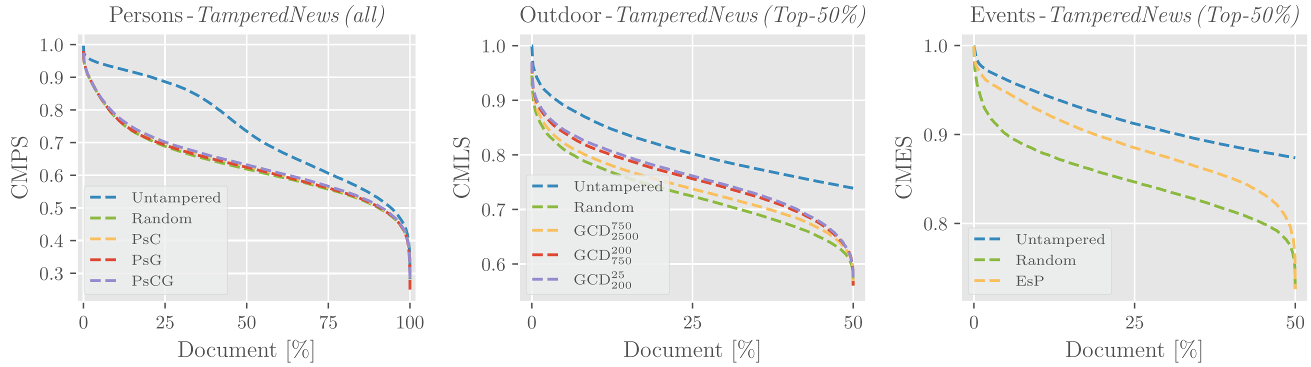


Fig. 4 Cross-modal similarity values of all (or a subset of) *TamperedNews* documents sorted in descending order for person, location (outdoor), and event entities and different tampering techniques (notations according to Sect. 5.1.1) (Color figure online)

rank all $2 \cdot |\mathbb{D}|$ documents. As suggested by previous work [21,36], the *Area Under Receiver Operating Curve* (AUC) is used for evaluation. We also propose to calculate the *Average Precision* (AP) for retrieving untampered (AP-clean) or tampered (AP-tampered) documents at specific recall levels R according to Eq. (4). In this respect, TP^i is the number of relevant documents at position i. For example, AP-tampered@25% describes the average precision when $|\mathbb{D}_R| = 0.25 \cdot |\mathbb{D}|$ of all tampered documents are retrieved.

$$\text{AP@R} = \frac{1}{|\mathbb{D}_R|} \sum_{i=1}^{k} \frac{\text{TP}^i}{i} \ , \tag{4}$$

Table 2 AUC for different operators Ψ (AC - agglomerative face clustering, Q_n - $n\%$ similarity quantile, max - max similarity) to calculate the cross-modal similarity for each entity of a given type (Sect. 3.3) within a document

Test set	$\lvert\mathbb{D}\rvert$	AC	Q_{75}	Q_{90}	Q_{95}	max
Persons: PsCG	16,848	0.93	0.92	**0.94**	**0.94**	0.90
Loc.-Outdoor: GCD$^{250}_{25}$	14,113	–	0.71	0.73	0.74	**0.77**
Loc.-Indoor: GCD$^{250}_{25}$	19,129	–	0.64	0.66	0.67	**0.69**
Events: EsP	7734	–	0.72	0.73	0.74	**0.75**

Results are reported for the amount of $|\mathbb{D}|$ documents in the respective *TamperedNews* (*Top*-50%) dataset with the hardest tampering strategy (notations according to Sect. 5.1.1)
Bold numbers indicate the best results for the individual test subsets

5.2.3 Test document selection for TamperedNews

Although the large size of the *TamperedNews* dataset allows for a large-scale analysis of the results, unfortunately a manual verification of cross-modal relations as for *News400* is infeasible. Thus, reporting the proposed metrics for the whole dataset can be misleading since it turned out during the annotation of *News400* that only a fraction of the documents has cross-modal entity correlations (Table 1). As discussed at the beginning of Sect. 3.3, it is possible that not a single entity mentioned in a news text is depicted in the corresponding image. To address this issue, we suggest measuring the metrics for specific subsets. More specifically, we consider the top-25% and top-50% documents (denoted as *TamperedNews (Top-k%)*) with respect to their cross-modal similarity of untampered entities since they more likely contain relations between image and text. This selection is also supported by the *Cross-modal Person Similarity* (CMPS) values for person verification (Fig. 4), which decrease more significantly after 25−50% of all documents and correspond to the percentage of manually verified documents in the *News400* dataset.

Please note, that experiments on top-$k\%$ subsets limit the comparability between two approaches to some degree. Depending on the specified parameters (e.g., feature descriptor, operator, etc.), the top-$k\%$ subsets comprise different documents. However, in Sect. 5.5 we explain how a meaningful comparison between two different approaches can be conducted.

5.3 Parameter selection

Face Clustering Threshold: The threshold $\tau_{\mathbb{P}}$ impacts the agglomerative clustering approach that filters retrieved face candidates for a person as explained in Sect. 3.3.1. For this reason, we have tested the *FaceNet* model [37] on the *Labeled Faces in the Wild* [20] benchmark and evaluated an optimal cosine similarity (normalized to the interval [0, 1]) threshold of $\tau_{\mathbb{P}} = 0.65$.

Operator for Cross-modal Similarities: In Sect. 3.3 we mentioned a number of possible operators Ψ such as the $n\%$-quantile or maximum to compute cross-modal similarity value based on the comparisons of all reference images of a specific entity to the news image. The results for AUC for different operators using a maximum of $k = 10$ reference images and all image sources (*Google*, *Bing*, and *Wikidata*)

Table 3 AUC using different image sources (W - Wikidata, G - Google, B - Bing) and maximum number of k images on the respective *TamperedNews* (*Top*-50%) subsets

Source	#Images			Persons	Locations		Events
					Outdoor	Indoor	
	k_W	k_G	k_B	PsCG	GCD_{200}^{25}	GCD_{200}^{25}	EsP
Google	–	20	–	**0.95**	0.76	0.68	0.73
Bing	–	–	20	0.90	0.76	**0.71**	**0.77**
All-10	all	10	10	0.93	0.77	0.69	0.75
All-20	all	20	20	0.93	**0.78**	**0.71**	0.76

Results are reported for the hardest tampering strategy (notations according to Sect. 5.1.1)

Bold numbers indicate the best results for the individual test subsets

on the respective *TamperedNews (Top-50%)* subsets are presented in Table 2.

For comparison, we also tested the face verification using the approach applied for event and location entities described in Sect. 3.3.2. Surprisingly, results for 90% and 95% quantiles are on par with the proposed person clustering. Also, contrary to our assumption that a quantile is more robust against noise for locations and events, it turned out that the maximum operator provides slightly better results for these entity types. This indicates that incorrect examples in the reference data have no significant impact on the performance. Except for person entities, where reference faces can be very similar, we assume that irrelevant or unrelated reference images less likely matches the entity depicted in the news image. In the remainder of this paper, results for persons are reported using the clustering strategy because we still believe that this is more robust in many scenarios. For locations and events the maximum operator is applied.

Amount and Sources of Reference Images: In total, we collected a maximum of $k = 20$ images from the image search engines of *Google* and *Bing* as well as all k_W available images on *Wikidata* (mostly one *Wikimedia* image) for each entity recognized in the text. We have used multiple sources to prevent possible selection biases of a specific image source and investigated the performance for different image sources and number of images. Since *Wikidata* usually only provides a single or sometimes no image for the linked entities, we exclude it from the comparison. The results on the respective *TamperedNews (Top-50%)* subsets for the AUC metric using the hardest tampering strategies are presented in Table 3.

They demonstrate that the performance using a single or all image sources is very similar. Also, the results using $k = 10$ reference images are almost identical compared to the maximum of $k = 20$ images. Hence, for the rest of our experiments, we use all available image sources with a maximum of $k = 10$ images per source as this provides a good trade-off between performance and speed and prevents possible selection biases.

5.4 Experimental results

In this section, we present the baseline results of the proposed system for cross-modal consistency verification on the *TamperedNews* (Sect. 5.4.1) and $News400$ dataset (Sect. 5.4.2).

5.4.1 Results for TamperedNews

Qualitative and quantitative results are presented in Figs. 4, 5, and Table 4. Results for all *TamperedNews* documents as well as the top-25% subset allow similar conclusions and are reported as supplemental material[3].

Results for Person Entities: As expected, person verification achieves the best performance since the entities and the retrieved example material are very unambiguous and neural networks for face recognition, such as *FaceNet* [37], can achieve impressive results. Despite the more challenging tampering techniques, our approach is still able to produce similar results. We have only experienced problems if persons were depicted in challenging conditions (e.g., extreme poses as shown in Fig. 5a for *John Kerry*) or were rather unknown, which results in false entity linking results and confusion with other persons (e.g., with a similar name).

Results for Location Entities: To evaluate performance for location entities, we distinguished between images of indoor and outdoor scenes using the scene probabilities $\hat{\boldsymbol{y}}_S$ extracted according to Sect. 4.3 and the hierarchy provided by the *Places365* dataset [47]. Due to the data diversity and ambiguity and the unequal distribution of photographs on earth, geolocation estimation is a complex problem that has attracted attention only in recent years [29,41,42]. Therefore, the results were expected to be worse compared to the person verification. Despite the complexity, good results were achieved for outdoor images, whereas the detection of modified indoor scenes is more challenging given the low amount of geographical cues and their ambiguity. However, even when entities are tampered with locations of similar appearance and low Great Circle Distance (GCD) (Fig. 5b, d), the system can operate on a good level and shows promising results.

In contrast to person entities, location entities are an instance of various parent classes such as *countries* or *cities*. For a more in depth-analysis, we have calculated the results for all types of locations separately using the documents $\mathbb{D}_s$ where an instance of a given type has achieved the highest *Cross-modal Locations Similarity* (CMLS) within the untampered set of entities. The results for some location types are presented in Table 5 (top) and show that the performance is best for more fine-grained entities such as *tourist attractions*, *buildings*, and *cities*. The performance for coarse location types such as *oceans*, *mountain ranges*, and *country states* are typically worse since they do not provide sufficient geographical cues or are too broad to retrieve suitable reference

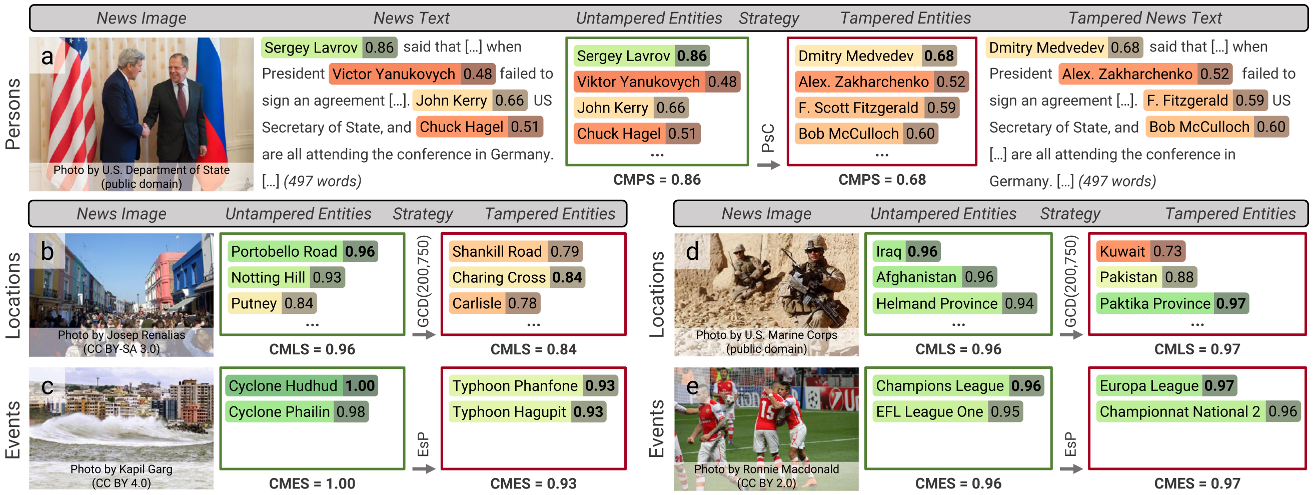


Fig. 5 Positive (**a**–**c**, *Cross-modal Similarity* of the untampered entity set is higher) and negative (**d**–**e**, *Cross-modal Similarity* of the tampered entity set is higher) verification results of some *TamperedNews* documents. Within each example the similarities (from red to green with intervals: persons [0.45, 1], locations [0.7, 1], events [0.8, 1]) of the news image to a set of untampered entities (green border) and tampered entities (red border) using *one* specific tampering strategy are shown. Images are replaced with similar ones due to license restrictions. Original images and full text are linked on the *GitHub* page[3] (Color figure online)

Table 4 Results for document verification (DV) and collection retrieval for the *TamperedNews (Top-50%)* dataset for different entity test sets

Test set	DV	Collection Retrieval						
	VA	AUC	AP-clean@			AP-tampered@		
			25%	50%	100%	25%	50%	100%
Persons (16848 documents)								
Random	0.94	0.95	96.08	95.45	92.64	100.0	100.0	96.16
PsC	0.93	0.94	95.53	94.67	91.68	100.0	100.0	95.61
PsG	0.94	0.95	95.77	95.07	92.27	100.0	100.0	96.00
PsCG	0.93	0.94	95.04	94.70	91.70	100.0	100.0	95.56
Locations-Outdoor (14113 documents)								
Random	0.88	0.85	92.57	88.02	81.71	100.0	100.0	88.82
GCD_{750}^{2500}	0.86	0.81	88.04	83.65	77.25	100.0	100.0	85.45
GCD_{200}^{750}	0.79	0.74	82.85	76.96	70.56	100.0	96.98	79.38
GCD_{25}^{200}	0.77	0.72	80.50	74.23	68.30	100.0	95.19	77.42
Locations-Indoor (19129 documents)								
Random	0.74	0.72	68.47	66.53	65.34	100.0	99.01	79.62
GCD_{750}^{2500}	0.73	0.70	63.62	62.86	62.72	100.0	97.57	77.80
GCD_{200}^{750}	0.74	0.71	66.93	65.10	63.97	100.0	97.70	78.14
GCD_{25}^{200}	0.69	0.68	55.99	57.74	59.48	100.0	95.97	76.04
Events (7734 documents)								
Random	0.92	0.91	92.20	91.26	87.61	100.0	100.0	93.66
EsP	0.75	0.71	70.72	67.30	64.92	100.0	96.72	77.68
Context (36217 documents)								
Random	0.81	0.80	88.95	83.03	76.32	100.0	100.0	84.79
top-25%	0.78	0.77	83.52	78.12	72.43	100.0	99.70	82.25
top-10%	0.76	0.74	77.76	73.21	68.78	100.0	98.33	79.84
top-5%	0.74	0.71	74.31	69.89	66.22	100.0	96.84	77.92

(notations according to Sect. 5.1.1)

Table 5 AUC for a selection of location (left) and event (right) types of the *TamperedNews* ($Top-50\%$) subset

Selection of 12 / 1,063 location entity types							Selection of 12 / 479 event entity types				
Type (#entities)	$\|\mathbb{D}\|$	$\|\mathbb{D}_s\|$	AUC				Type (#entities)	$\|D\|$	$\|D_s\|$	AUC	
			Random	GCD_{750}^{2500}	GCD_{200}^{750}	GCD_{25}^{200}				Random	EsP
Continent (7)	1510	116	0.82	0.76	0.77	0.78	Football clubs cup (3)	1094	801	0.93	0.49
Country (184)	9516	2892	0.86	0.76	0.72	0.70	Sport competition (14)	155	111	0.95	0.76
State (109)	1969	411	0.87	0.75	0.73	0.71	Festival (64)	516	421	0.90	0.77
City (706)	9781	3333	0.86	0.83	0.78	0.75	Award (6)	260	206	0.91	0.74
Town (592)	4774	1655	0.80	0.88	0.71	0.69	Holiday (28)	285	141	0.91	0.84
Street (24)	300	64	0.80	0.78	0.76	0.73	Television series (16)	144	123	0.87	0.70
Tourist attraction (63)	880	172	0.94	0.91	0.90	0.88	War (39)	919	665	0.83	0.67
Building (40)	640	77	0.91	0.87	0.90	0.88	Murder (19)	154	134	0.93	0.78
Mountain range (13)	201	42	0.94	0.86	0.70	0.62	Disaster (5)	70	61	0.93	0.91
Mountain (9)	85	31	0.97	0.92	0.79	0.77	Scandal (10)	112	95	0.95	0.68
Ocean (4)	344	59	0.89	0.63	0.58	0.63	Protest (7)	60	54	0.89	0.67
River (36)	412	106	0.90	0.82	0.78	0.71	Legal case (10)	37	34	0.89	0.79

$|\mathbb{D}|$ is the number of documents containing at least one entity of this type and $|\mathbb{D}_s|$ is the number of times this type has achieved the highest cross-modal similarity within the untampered set. Results are reported for the documents $\mathbb{D}_s$ of each entity type

images. Although the results for *continents* or *countries* are also comparatively high, we believe the reason is that the candidates for tampering are easier to distinguish since locations of those types have higher geographical and cultural differences. The tampering is much more challenging for fine-grained entities, as illustrated in Fig. 5b, d.

Results for Event Entities: In general, good results were achieved for event verification. As for locations we have provided results of common event types in Table 5 (bottom). While the results for *festivals*, *holiday*, and *disasters* are promising, event types such as *football club competitions*, *protests*, and *wars* are hard to distinguish. We believe that this is caused by the high visual similarity of events within these types. For example, many news articles on *football clubs cups* contain images which, unlike articles on *sport competitions* that refer to different types of sports, depict typical scenes (e.g, players on the pitch) of the same sport. Thus, reference images for the different competitions are very similar. Moreover, the utilized event classification approach [31] distinguishes between event types such as *football*, *elections*, or types of *natural disasters* rather than between sub-types or concrete event instances such as *UEFA Champions League* or *2020 U.S. elections*. Despite these limitations the results are superior to the scene classification approach used in our previous work [30], as discussed in more detail in Sect. 5.5.

Cross-modal Context Similarity: The results for scene context verification indicate that our system can reliably detect documents with randomly changed images. However, as also stated by [36], this task is rather easy as the semantic relations are not maintained. When similar images are used for tampering, this task becomes much more challenging. Since networks for object classification (used for tampering) and scene classification (used for verification) can produce comparable results, performance is steadily decreasing using more similar images for tampering that might even show the same scene, e.g., *sport*. However, our system is still able to hint towards cross-modal consistencies.

5.4.2 Results for News400

Since the number of documents is rather limited and the cross-modal mutual presence of entities was manually verified, results for *News400* are reported for all documents with verified relations. Based on the results displayed in Table 6, similar conclusions on the overall system performance can be drawn. However, results while retrieving tampered documents are noticeably worse. This is mainly caused by the fact that some untampered entities with valid cross-modal relations can be either unspecific (e.g., mentioning of a *country*) or the retrieved images for visual verification do not fit the document's image content. Since subsets of the top-$k\%$ documents for *TamperedNews* were used to counteract the influence of untampered documents that do not show any

Table 6 Results for document verification (DV) and collection retrieval for the *News400* dataset

Test set	DV	Collection Retrieval						
	VA	AUC	AP-clean@			AP-tampered@		
			25%	50%	100%	25%	50%	100%
Persons (116 verified documents)								
Random	0.95	0.91	100.0	100.0	93.70	85.19	85.96	85.95
PsC	0.92	0.90	100.0	99.49	92.14	83.07	84.52	84.14
PsG	0.91	0.90	99.10	98.40	92.34	82.40	84.36	84.64
PsCG	0.92	0.91	100.0	100.0	94.00	84.38	85.25	85.60
Locations-Outdoor (54 verified documents)								
Random	0.89	0.85	100.0	98.01	87.72	83.19	80.76	79.47
GCD_{750}^{2500}	0.81	0.80	92.61	89.94	81.49	72.19	72.95	73.20
GCD_{200}^{750}	0.80	0.74	86.70	82.42	74.76	63.03	66.73	67.33
GCD_{25}^{200}	0.80	0.72	86.70	82.25	72.85	63.59	67.97	66.35
Locations-Indoor (15 verified documents)								
Random	0.80	0.75	91.67	80.94	74.85	88.75	86.44	77.20
GCD_{750}^{2500}	0.67	0.64	62.20	58.74	60.04	80.42	82.28	69.37
GCD_{200}^{750}	0.87	0.69	85.42	74.31	68.90	88.75	85.23	72.96
GCD_{25}^{200}	0.80	0.62	69.17	62.64	61.40	80.42	78.06	67.12
Events (31 verified documents)								
Random	1.00	0.93	100.0	96.18	92.58	100.0	99.63	93.93
EsP	0.74	0.72	63.44	66.19	65.49	89.57	86.45	74.76
Context (91 verified documents)								
Random	0.70	0.70	87.03	87.50	73.62	61.11	63.09	63.19
top-25%	0.70	0.68	92.19	88.43	72.96	53.60	57.77	59.69
top-10%	0.64	0.66	70.54	74.12	65.58	56.15	59.72	59.75
top-5%	0.66	0.63	74.48	73.09	64.18	50.77	55.99	56.98

Results are reported for all verified documents for different entity test sets (notations according to Sect. 5.1.1)

cross-modal relations (as discussed in Sect. 5.2.3) this problem was bypassed. We have verified the same behavior for *News400* when experimenting on these subsets. For more details, we refer to the supplemental material[3]. In addition, performance for context verification is worse compared to *TamperedNews*. We assume that this is due to the less powerful word embedding for the *German* language. Overall, the system achieves promising performance for cross-modal consistency verification. Since it dynamically gathers example data from the Web, it is robust to changes in topics and entities, even when applied to news articles from another country and publication date.

5.5 Comparison of event feature descriptors

As discussed in Sect. 3.2, we decided to use the ontology-driven event classification approach [31] to compute event features for our proposed system. Due to the absence of suitable methods for event classification, a more general scene classification model was applied in our previous approach [30]. It is trained on 365 places covered by the *Places365* dataset [47] and the visual features $\mathbf{f}_E$ are obtained from the last pooling layer of a *ResNet-50* model[8] [14,15].

To compare both approaches, we evaluate their performances on the *News400* dataset as it contains documents with verified event relations. As explained in Sect. 5.2.3 we have used the *TamperedNews* ($Top - 50\%$) documents as subsets for testing since they more likely contain cross-modal relations. However, this complicates the comparison of two approaches as those subsets can be different depending on their specified parameters (feature descriptor, operator, etc.). Thus, we report results on all documents as well as on the intersection and union of the *TamperedNews* ($Top - 50\%$) document sets of both approaches. In this way, the test sets contain documents that are either considered relevant from both or at least one approach, respectively. The results are presented in Table 7 and demonstrate that the event classification approach achieves superior performances. However, as already discussed in Sect. 5.4.1 the approach is not trained for the classification of concrete event instances and instead focuses on more generic event types. As a consequence, improvements for the EsP test set containing tampered events

Table 7 *Verification Accuracy* (VA) and *Area under Receiver Operating Curve* using a scene classification network trained on *Places365* [47] (denoted as ”Scene”) and an ontology-driven event classification approach trained on *VisE-D* [31] (denoted as ”Event”) as feature extractor for different event tampering strategies and datasets

Method	Event Tampering Technique			
	Random		EsP	
	VA	**AUC**	**VA**	**AUC**
News400 (31 verified documents)				
Scene-CNN [47]	0.87	0.85	0.68	0.64
Event-CNN [31]	**1.00**	**0.93**	**0.74**	**0.72**
TamperedNews (all 15,467 documents)				
Scene-CNN [47]	0.67	0.66	0.59	0.56
Event-CNN [31]	**0.70**	**0.70**	**0.59**	**0.57**
TamperedNews (Top-50% intersection—6080 documents)				
Scene-CNN [47]	0.91	0.89	0.75	0.70
Event-CNN [31]	**0.94**	**0.93**	**0.76**	**0.71**
TamperedNews (Top-50% union—9388 documents)				
Scene-CNN [47]	0.83	0.82	0.70	0.65
Event-CNN [31]	**0.86**	**0.86**	**0.71**	**0.67**

Bold numbers indicate the best results for the individual test subsets

of the same parent class are not as significant as for the randomly tampered test set. Further limitations and dependencies are discussed in the next section.

5.6 Limitations and dependencies

News covered in the World Wide Web are dynamic and new entities and topics evolve every day. We have deliberately chosen *Wikifier* for named entity linking as it can dynamically cover *Wikipedia* entities. However, the proposed system is restricted to entities that exist in a knowledge base. Besides, the system relies on the rankings and response times of image search engines. In this regard, the reference images for coarse entities such as *countries* or *continents* crawled from the Web might not match the news image. Some named entities such as ”Hanover” (German or U.S. city) or ”Tesla” (company or inventor) can also be ambiguous. Referring to Fig. 1, we also noticed that querying entities such as the city ”Liverpool” retrieves images that depict another more popular entity, in this case the football club ”Liverpool F.C.” rather than the actual entity.

A potential solution to the aforementioned problems is to include knowledge graph information and relations that are already extracted by the system. For example, adding the country (*Wikidata* property *P17*) ”Germany” to the query ”Hanover” (*Wikidata* item *Q1715*) or using the entity type (*Wikidata* property *P31*) ”city” in combination with the query ”Liverpool” (*Wikidata* item *Q24826*) can prevent potential ambiguities.

6 Conclusions

In this paper, we have presented a novel analytics system and benchmark datasets to measure the cross-modal consistency in real-world news articles. Named entity linking is applied to find persons, locations, and events in the news text. Reference data is automatically gathered from the Web and used in combination with novel measures of cross-modal similarity for the visual verification of entities in the article’s photograph. In this regard, state-of-the-art computer vision methods are applied. Furthermore, a more general measure of cross-modal similarity of the textual content to the scene depicted in the image has been introduced. Unlike previous work, our system is completely unsupervised and visual representations of the extracted entities are not derived from similar data sources with additionally available metadata. Experiments were conducted on two datasets that contain real-world news articles across different topics, domains, and languages and have clearly demonstrated the feasibility of the proposed approach.

As mentioned in Sect. 5.6 the system performance for coarse (*countries*, *continents*, etc.), ambiguous, or less popular entities can suffer due to the lack of relevant reference images crawled by the unsupervised Web image search. Thus, we aim to refine the image search queries based on the extracted named entities for the visual verification approach by further exploiting knowledge graph information and entity relationships in the future. Furthermore, the event classification approach is only able to distinguish between *event types* such as types of *sports*, *natural disasters*, *elections*, etc. The system can greatly benefit from an event classification approach that is capable of differentiating between more fine-grained event types and concrete *event instances*, e.g., *UEFA Champions League*, *2020 U.S. elections*, etc. Another interesting direction of research is to investigate the impact of other types of entities such as time or organizations, entity relations, as well as relations between the overall textual and visual sentiment.

Acknowledgements This work has partially received funding from the European Union’s Horizon research and innovation programme 2020 under the Marie Skłodowska-Curie Grant Agreement No 812997, and the German Research Foundation (DFG: Deutsche Forschungsgemeinschaft, project number: 388420599). We are very grateful to Avishek Anand (L3S Research Center, Leibniz University Hannover) for his valuable comments that improved the quality of the paper.

Funding Open Access funding enabled and organized by Projekt DEAL.

Open Access This article is licensed under a Creative Commons Attribution 4.0 International License, which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons licence, and indicate if changes were made. The images or other third party material in this article are included in the article's Creative Commons licence, unless indicated otherwise in a credit line to the material. If material is not included in the article's Creative Commons licence and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder. To view a copy of this licence, visit http://creativecommons.org/licenses/by/4.0/.

References

1. Ahmad K, Conci N, Boato G, Natale FGBD (2016) USED: a large-scale social event detection dataset. In: Timmerer C (ed) Proceedings of the 7th international conference on multimedia systems, MMSys 2016, Klagenfurt, Austria, May 10–13, 2016, pp 50:1–50:6. ACM. https://doi.org/10.1145/2910017.2910624
2. Ahsan U, Sun C, Hays J, Essa IA (2017) Complex event recognition from images with few training examples. In: 2017 IEEE winter conference on applications of computer vision, WACV 2017, Santa Rosa, CA, USA, March 24–31, 2017, pp 669–678. IEEE Computer Society. https://doi.org/10.1109/WACV.2017.80
3. Barthes R (1977) Image-music-text, ed. and trans. S. Heath, London: Fontana, 332
4. Bateman J (2014) Text and image: a critical introduction to the visual/verbal divide. Routledge, Milton Park
5. Brank J, Leban G, and Grobelnik M (2018) Semantic annotation of documents based on wikipedia concepts. Informatica (Slovenia), 42(1), http://www.informatica.si/index.php/informatica/article/view/2228
6. Broersma M, Graham T (2013) Twitter as a news source: how Dutch and British newspapers used tweets in their news coverage, 2007–2011. J Pract 7(4):446–464. https://doi.org/10.1080/17512786.2013.802481
7. Chen B, Ghosh P, Morariu VI, Davis LS (2017) Detection of metadata tampering through discrepancy between image content and metadata using multi-task deep learning. In: 2017 IEEE conference on computer vision and pattern recognition workshops, CVPR Workshops 2017, Honolulu, HI, USA, July 21–26, 2017, pp 1872–1880. IEEE Computer Society. https://doi.org/10.1109/CVPRW.2017.234
8. Chen B-C, Davis LS (2019) Deep representation learning for metadata verification. In: 2019 IEEE winter applications of computer vision workshops (WACVW), pp 73–82. IEEE
9. Deng J, Dong W, Socher R, Li L, Li K, Li F (2009) Imagenet: a large-scale hierarchical image database. In: 2009 IEEE computer society conference on computer vision and pattern recognition (CVPR 2009), 20–25 June 2009, Miami, Florida, USA, pp 248–255. IEEE Computer Society. https://doi.org/10.1109/CVPR.2009.5206848
10. Gottschalk S, Demidova E (2018) Eventkg: a multilingual event-centric temporal knowledge graph. In: Gangemi A, Navigli R, Vidal M, Hitzler P, Troncy R, Hollink L, Tordai A, Alam M (eds) The semantic web–15th international conference, ESWC 2018, Heraklion, Crete, Greece, June 3–7, 2018, Proceedings, volume 10843 of Lecture Notes in Computer Science, pp 272–287. Springer. https://doi.org/10.1007/978-3-319-93417-4_18
11. Gottschalk S, Demidova E (2019) Eventkg—the hub of event knowledge on the web—and biographical timeline generation. Semant Web 10(6):1039–1070. https://doi.org/10.3233/SW-190355
12. Grave E, Bojanowski P, Gupta P, Joulin A, Mikolov T (2018) Learning word vectors for 157 languages. In: Calzolari N, Choukri K, Cieri C, Declerck T, Goggi S, Hasida K, Isahara H, Maegaard B, Mariani J, Mazo H, Moreno A, Odijk J, Piperidis S, Tokunaga T (eds) Proceedings of the eleventh international conference on language resources and evaluation, LREC 2018, Miyazaki, Japan, May 7–12, 2018. European Language Resources Association (ELRA). http://www.lrec-conf.org/proceedings/lrec2018/summaries/627.html
13. Halliday MAK, Matthiessen CM (2013) Halliday's introduction to functional grammar. Routledge, Milton Park. https://doi.org/10.4324/9780203431269
14. He K, Zhang X, Ren S, Sun J (2016a) Deep residual learning for image recognition. In: 2016 IEEE conference on computer vision and pattern recognition, CVPR 2016, Las Vegas, NV, USA, June 27–30, 2016, pp 770–778. IEEE Computer Society. https://doi.org/10.1109/CVPR.2016.90
15. He K, Zhang X, Ren S, Sun J (2016b) Identity mappings in deep residual networks. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016—14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part IV, volume 9908 of Lecture Notes in Computer Science, pp 630–645. Springer. https://doi.org/10.1007/978-3-319-46493-0_38
16. Henning CA, and Ewerth R (2017) Estimating the information gap between textual and visual representations. In: Ionescu B, Sebe N, Feng J, Larson MA, Lienhart R, Snoek C (eds) Proceedings of the 2017 ACM on international conference on multimedia retrieval, ICMR 2017, Bucharest, Romania, June 6–9, 2017, pp 14–22. ACM. https://doi.org/10.1145/3078971.3078991
17. Henning CA, Ewerth R (2018) Estimating the information gap between textual and visual representations. Int J Multim Inf Retr 7(1):43–56. https://doi.org/10.1007/s13735-017-0142-y
18. Hoffart J, Yosef MA, Bordino I, Fürstenau H, Pinkal M, Spaniol M, Taneva B, Thater S, Weikum G (2011) Robust disambiguation of named entities in text. In: Proceedings of the 2011 conference on empirical methods in natural language processing, EMNLP 2011, 27–31 July 2011, John McIntyre Conference Centre, Edinburgh, UK, A meeting of SIGDAT, a Special Interest Group of the ACL, pp 782–792. ACL. https://www.aclweb.org/anthology/D11-1072/
19. Honnibal M and Montani I (2017) spaCy 2: natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. To appear
20. Huang GB, Ramesh M, Berg T, Learned-Miller E (2007) Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical Report 07-49, University of Massachusetts, Amherst, 10
21. Jaiswal A, Sabir E, Abd-Almageed W, Natarajan P (2017) Multimedia semantic integrity assessment using joint embedding of images and text. In: Liu Q, Lienhart R, Wang H, Chen SK, Boll S, Chen YP, Friedland G, Li J, Yan S (eds) Proceedings of the 2017 ACM on multimedia conference, MM 2017, Mountain View, CA, USA, October 23–27, 2017, pp 1465–1471. ACM. https://doi.org/10.1145/3123266.3123385
22. Jaiswal A, Wu Y, AbdAlmageed W, Masi I, Natarajan P (2019) AIRD: adversarial learning framework for image repurposing detection. In: IEEE conference on computer vision and pattern recognition, CVPR 2019, Long Beach, CA, USA, June 16–20, 2019, pp 11330–11339. Computer Vision Foundation / IEEE. https://doi.org/10.1109/CVPR.2019.01159. http://openaccess.thecvf.com/content_CVPR_2019/html/Jaiswal_AIRD_Adversarial_Learning_Framework_for_Image_Repurposing_Detection_CVPR_2019_paper.html
23. Kakar P, Sudha N (2012) Verifying temporal data in geotagged images via sun azimuth estimation. IEEE Trans Inf Forensics Secur 7(3):1029–1039. https://doi.org/10.1109/TIFS.2012.2188796
24. Kolitsas N, Ganea O, Hofmann T (2018) End-to-end neural entity linking. In: Korhonen A, Titov I (eds) Proceedings of the 22nd conference on computational natural language learning, CoNLL 2018, Brussels, Belgium, October 31—November 1, 2018, pp 519–529. Association for Computational Linguistics. https://doi.org/10.18653/v1/k18-1050

25. Kruk J, Lubin J, Sikka K, Lin X, Jurafsky D, Divakaran A (2019) Integrating text and image: determining multimodal document intent in instagram posts. In: Inui K, Jiang J, Ng V, Wan X (eds) Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3–7, 2019, pp 4621–4631. Association for Computational Linguistics. https://doi.org/10.18653/v1/D19-1469
26. Li X, Xu W, Wang S, Qu X (2017) Are you lying: validating the time-location of outdoor images. In: Gollmann D, Miyaji A, Kikuchi H (eds) Applied cryptography and network security–15th international conference, ACNS 2017, Kanazawa, Japan, July 10–12, 2017, Proceedings, volume 10355 of Lecture Notes in Computer Science, pp 103–123. Springer. https://doi.org/10.1007/978-3-319-61204-1_6
27. Marsh EE, White MD (2003) A taxonomy of relationships between images and text. J Doc 59(6):647–672. https://doi.org/10.1108/00220410310506303
28. Martinec R, Salway A (2005) A system for image-text relations in new (and old) media. Vis Commun 4(3):337–371. https://doi.org/10.1177/1470357205055928
29. Müller-Budack E, Pustu-Iren K, Ewerth R (2018) Geolocation estimation of photos using a hierarchical model and scene classification. In: Ferrari V, Hebert M, Sminchisescu C, Weiss Y (eds) Computer vision—ECCV 2018—15th European conference, Munich, Germany, September 8–14, 2018, Proceedings, Part XII, volume 11216 of Lecture Notes in Computer Science, pp 575–592. Springer. https://doi.org/10.1007/978-3-030-01258-8_35
30. Müller-Budack E, Theiner J, Diering S, Idahl M, Ewerth R (2020) Multimodal analytics for real-world news using measures of cross-modal entity consistency. In: Gurrin C, Jónsson BT, Kando N, Schöffmann K, Chen YP, O'Connor NE (eds), Proceedings of the 2020 on international conference on multimedia retrieval, ICMR 2020, Dublin, Ireland, June 8–11, 2020, pp 16–25. ACM. https://doi.org/10.1145/3372278.3390670
31. Müller-Budack E, Springstein M, Hakimov S, Mrutzek K, Ewerth R (2021) Ontology-driven event type classification in images. In: Proceedings of the IEEE/CVF winter conference on applications of computer vision, pp 2928–2938, 2021
32. Otto C, Springstein M, Anand A, Ewerth R (2019) Understanding, categorizing and predicting semantic image-text relations. In: El-Saddik A, Bimbo AD, Zhang Z, Hauptmann AG, Candan KS, Bertini M, Xie L, Wei X (eds) Proceedings of the 2019 on international conference on multimedia retrieval, ICMR 2019, Ottawa, ON, Canada, June 10–13, 2019, pp 168–176. ACM, 2019. https://doi.org/10.1145/3323873.3325049
33. Ramisa A, Yan F, Moreno-Noguer F, Mikolajczyk K (2018) Breakingnews: article annotation by image and text processing. IEEE Trans Pattern Anal Mach Intell 40(5):1072–1085. https://doi.org/10.1109/TPAMI.2017.2721945
34. Rizzo G, Troncy R (2012) NERD: a framework for unifying named entity recognition and disambiguation extraction tools. In: Daelemans W, Lapata M, Màrquez L (eds) EACL 2012, 13th conference of the european chapter of the association for computational linguistics, Avignon, France, April 23–27, 2012, pp 73–76. The Association for Computer Linguistics. https://www.aclweb.org/anthology/E12-2015/
35. Rogers R (2013) Debanalizing twitter: the transformation of an object of study. In: Davis HC, Halpin H, Pentland A, Bernstein M, Adamic LA (eds) Web science 2013 (co-located with ECRC), WebSci '13, Paris, France, May 2–4, 2013, pp 356–365. ACM, 2013. https://doi.org/10.1145/2464464.2464511
36. Sabir E, AbdAlmageed W, Wu Y, Natarajan P (2018) Deep multimodal image-repurposing detection. In: Boll S, Lee KM, Luo J, Zhu W, Byun H, Chen CW, Lienhart R, Mei T (eds) 2018 ACM multimedia conference on multimedia conference, MM 2018, Seoul, Republic of Korea, October 22–26, 2018, pp 1337–1345. ACM. https://doi.org/10.1145/3240508.3240707
37. Schroff F, Kalenichenko D, Philbin J (2015) Facenet: a unified embedding for face recognition and clustering. In: IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pp 815–823. IEEE Computer Society, 2015. https://doi.org/10.1109/CVPR.2015.7298682
38. Smeulders AWM, Worring M, Santini S, Gupta A, Jain RC (2000) Content-based image retrieval at the end of the early years. IEEE Trans Pattern Anal Mach Intell 22(12):1349–1380. https://doi.org/10.1109/34.895972
39. Tandoc EC Jr, Johnson E (2016) Most students get breaking news first from twitter. Newsp Res J 37(2):153–166. https://doi.org/10.1177/0739532916648961
40. Unsworth L (2007) Image/text relations and intersemiosis: towards multimodal text description for multiliteracies education. In: Proceedings of the 33rd international systemic functional congress, pp 1165–1205
41. Vo NN, Jacobs N, Hays J (2017) Revisiting IM2GPS in the deep learning era. In: IEEE international conference on computer vision, ICCV 2017, Venice, Italy, October 22–29, 2017, pp 2640–2649. IEEE Computer Society. https://doi.org/10.1109/ICCV.2017.286
42. Weyand T, Kostrikov I, Philbin J (2016) Planet—photo geolocation with convolutional neural networks. In: Leibe B, Matas J, Sebe N, Welling M (eds) Computer vision—ECCV 2016—14th European conference, Amsterdam, The Netherlands, October 11–14, 2016, Proceedings, Part VIII, volume 9912 of Lecture Notes in Computer Science, pp 37–55. Springer, 2016. https://doi.org/10.1007/978-3-319-46484-8_3
43. Xiong Y, Zhu K, Lin D, Tang X (2015) Recognize complex events from static images by fusing deep channels. In: IEEE conference on computer vision and pattern recognition, CVPR 2015, Boston, MA, USA, June 7–12, 2015, pp 1600–1609. IEEE Computer Society. https://doi.org/10.1109/CVPR.2015.7298768
44. Ye K, Honarvar Nazari N, Hahn J, Hussain Z, Zhang M, Kovashka A (2019) Interpreting the rhetoric of visual advertisements. IEEE Trans Pattern Anal Mach Intell, pp 1–1, 2019. ISSN 1939-3539. https://doi.org/10.1109/TPAMI.2019.2947440
45. Zhang K, Zhang Z, Li Z, Qiao Y (2016) Joint face detection and alignment using multitask cascaded convolutional networks. IEEE Signal Process Lett 23(10):1499–1503. https://doi.org/10.1109/LSP.2016.2603342
46. Zhang M, R Hwa R, and Kovashka A (2018) Equal but not the same: understanding the implicit relationship between persuasive images and text. In: British machine vision conference 2018, BMVC 2018, Newcastle, UK, September 3–6, 2018, p 8. BMVA Press, 2018. http://bmvc2018.org/contents/papers/0228.pdf
47. Zhou B, Lapedriza À, Khosla A, Oliva A, Torralba A (2018) Places: a 10 million image database for scene recognition. IEEE Trans Pattern Anal Mach Intell 40(6):1452–1464. https://doi.org/10.1109/TPAMI.2017.2723009

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.