**Noname manuscript No.**
(will be inserted by the editor)

# Multilingual Visual Sentiment Concept Clustering and Analysis

**Nikolaos Pappas**[†] · **Miriam Redi**[†] · **Mercan Topkara**[†] · **Hongyi Liu**[†] · **Brendan Jou** · **Tao Chen** · **Shih-Fu Chang**

Received: date / Accepted: date

**Abstract** Visual content is a rich medium that can be used to communicate not only facts and events, but also emotions and opinions. In some cases, visual content may carry a universal affective bias (e.g., natural disasters or beautiful scenes). Often however, to achieve a parity in the affections a visual media invokes in its recipient compared to the one an author intended requires a deep understanding and even sharing of cultural backgrounds. In this study, we propose a computational framework for the clustering and analysis of multilingual visual affective concepts used in different languages which enable us to pinpoint alignable differences (via similar concepts) and non-alignable differences (via unique concepts) across cultures. To do so, we crowdsource sentiment labels for the MVSO dataset, which contains 16K multilingual visual sentiment concepts and 7.3M images tagged with these concepts. We then represent these concepts in a distribution-based word vector space via (1) pivotal translation or (2) cross-lingual semantic alignment. We then evaluate these representations on three tasks: affective concept retrieval, concept clustering, and sentiment prediction - all across languages. The proposed clustering framework enables the analysis of the large multilingual dataset both quantitatively and qualitatively. We also show a novel use case consisting of a facial image data subset and explore cultural insights about visual sentiment concepts in such portrait-focused images.

**Keywords** Multilingual · Language; Cultures; Cross-cultural · Emotion · Sentiment · Ontology · Concept Detection · Social Multimedia

[†] Denotes equal contribution.

Nikolaos Pappas
Idiap Research Institute, Martigny, Switzerland
E-mail: npappas@idiap.ch

Miriam Redi
Nokia Bell Labs, Cambridge, United Kingdom
E-mail: redi@belllabs.com

Mercan Topkara
Teachers Pay Teachers, New York, NY, USA
E-mail: mercan@teacherspayteachers.com

Brendan Jou · Hongyi Liu · Tao Chen · Shih-Fu Chang
Columbia University, New York, NY, USA
E-mail: {bjou, hongyi.liu, taochen, sfchang}@ee.columbia.edu

## 1 Introduction

Everyday, billions of users from around the world share their visual memories on online photo sharing platforms. Web users speak hundreds of different languages, come from different countries and backgrounds. Such multicultural diversity also results in users representing the visual world in very different ways. For instance, [1] showed that Flickr users with different cultural backgrounds use different concepts to describe visual emotions. But how can we build tools to analyze and retrieve multimedia data related to sentiments and emotions in visual content that arise from such influence of diverse cultural background? Multimedia retrieval in a multicultural environment cannot be independent of the language used by users to describe their visual content.

For example, in the vast sea of photo sharing content on platforms such as Flickr, it is easy to find pictures of traditional costumes from all around the world. However, a basic keyword search, e.g. *traditional costumes*, does not return rich multicultural results. Instead, returned content often comes from Western countries, especially from countries where English is the primary language. The problem we tackle is to analyze and develop a deeper understanding of multicultural content in the context of a large social photo sharing platform. A purely image-based analysis would not provide a complete understanding since it only cluster visually-similar images together, missing the differences between cultures, e.g. how an *old house* or *good food*

might look in each culture. We mitigate these problems of pure image-based analysis with the aid of computational language tools, and their combination with visual feature analysis.

This paper focuses on two dimensions characterizing users' cultural background: language and sentiment. Specifically, we aim to understand how do people textually describe sentiment concepts in their languages and how similar concepts or images may carry different degrees of sentiments in various languages. To the best of our knowledge, we have built the first complete framework for analyzing, exploring, and retrieving multilingual emotion-biased visual concepts to our knowledge. This allows us to retrieve examples of concepts such as *traditional costumes* from visual collections of different languages (see Fig. 1). To this end, we adopt the Multilingual Visual Concept Ontology (MVSO) dataset [1] to semantically understand and compare visual sentiment concepts across multiple languages. This allows us to investigate various aspects of the MVSO, including (1) visual differences for images related to similar visual concepts across languages and (2) cross-culture differences, by discovering visual concepts that are unique to each language.

To achieve this, it is essential to match lexical expressions of concepts from one language to another. One naïve solution is through *exact matching*, an approach where we translate of all languages to a single one as the pivot, e.g. English. However, given that lexical choices for the same concepts vary across languages, the exact matching of multilingual concepts has a small coverage across languages. To overcome this sparsity issue, we propose an *approximate matching* approach which represents multilingual concepts in a common semantic space based on pre-trained word embeddings via translation to a pivot language or through semantic alignment of monolingual embeddings. This allows us to compute the semantic proximity or distance between visual sentiment concepts and cluster concepts from multiple languages. Furthermore, it enables a better connectivity between visual sentiment concepts of different languages, and the discovery of multilingual clusters of visual sentiment concepts, whereas exact matching clusters are mostly dominated by a single language. The contributions of this paper can be summarized as follows:

1. We design a crowdsourcing process to annotate the sentiment score of visual concepts from 11 languages in MVSO, and thus create the largest publicly available labeled multilingual visual sentiment dataset for research in this area.
2. We evaluate and compare a variety of unsupervised distributed word and concept representations on visual concept matching. In addition, we define a novel evaluation metric called *visual semantic relatedness*.


**Fig. 1** Example images from four languages from the same cluster related to "traditional clothing" concept. Even though all images are tagged with semantically similar concepts, each culture interprets such concepts with different visual patterns and sentimental values.

3. We design new tools to evaluate sentiment and semantic consistency on various multilingual sentiment concept clustering results.
4. We evaluate the concept representations in several applications, including cross-language concept retrieval, sentiment prediction, and unique cluster discovery. Our results confirm the performance gains by fusing multimodal features.
5. We demonstrate the performance gain in sentiment prediction by fusing features from language and image modalities.
6. We perform a thorough qualitative analysis and a novel case study of portrait images in MVSO. We find that Eastern and Western languages tend to attach different sentiment concepts to portrait images, but all languages attach mostly positive concepts to face pictures.

This study extends our prior work in [35] by introducing a new multilingual concept sentiment prediction task (Section 7), comparing different concept representations over three distinct tasks (Sections 5, 6, 7), and performing an in-depth qualitative analysis with the goal of discovering interesting multilingual and monolingual clusters (Section 8). To highlight the novel insights discovered in each of our comprehensive studies, we will display the text about each insight in the bold font.

The rest of the paper is organized as follows: Section 2 discusses the related work; Section 3 describes our visual sentiment crowdsourcing results, while Section 4, describes approaches for matching visual sentiment concepts; the evaluation results on concept retrieval and clustering are analyzed in Sections 5 and 6 respectively, while the visual sentiment concept prediction resuls are in Section 7; Section 8 contains our qualitative analysis, and Section 9 describes a clustering case-study on portait images. Lastly, Section 10 concludes the paper and provides future directions.

## 2 Related Work

### 2.1 Visual Sentiment Analysis

In computational sentiment analysis, the goal is typically to detect the overall disposition of an individual, specifically as 'positive' or 'negative,' towards an object or event manifesting in some medium (digital or otherwise) [36,38,39,41–44], or to detect categorical dispositions such as the sentiment towards a stimulus' aspects or features [45–51]. While this research area had originally focused more on the linguistic modality, wherein text-based media are analyzed for opinions and sentiment, later it was extended to other modalities like visual and audio [52,53,55,54,57,56,59]. In particular, [52] addressed the problem of tri-modal sentiment analysis and showed that sentiment understanding can benefit from joint exploitation of all modalities. This was also confirmed in [53] on multimodal sentiment analysis study of Spanish videos. More recently, [57,59] improved over previous state-of-the-art using a deep convolutional network for utterance-level multimodal sentiment analysis. And in another line research, in bi-modal sentiment analysis, [55] proposed a large-scale visual sentiment ontology (VSO) and showed that using both visual and text features for predicting the sentiment of a tweet improves over individual modalities. Based on VSO, [1] proposed an even larger-scale multilingual visual sentiment ontology (MVSO), which analyzed the sentiment and emotions across twelve different languages and performed sentiment analysis on images. In the present study, instead of using automatic sentiment tools to detect the sentiment of a visual concept as in [55,1,35], we perform a large-scale human study in which we annotate the sentiment of visual concepts based on both visual and linguistic modalities, and, furthermore, we propose a new task for detecting the visual sentiment of adjective-noun-pairs based on its compound words and sample of images in which they are used as tags.

### 2.2 Distributed Word Representations

Research on distributed word representations [2–5] has recently extended to multiple languages either by using bilingual word alignments or parallel corpora to transfer linguistic information from multiple languages. For instance, [6] proposed to learn distributed representations of words across languages by using a multilingual corpus from Wikipedia. [7,8] proposed to learn bilingual embeddings in the context of neural language models utilizing multilingual word alignments. [9] proposed to learn joint-space embeddings across multiple languages without relying on word alignments. Similarly, [10] proposed auto-encoder-based methods to learn multilingual word embeddings. A limitation when dealing with many languages is the scarcity of data for all pairs. In the present study, we use a pivot language to align the multiple languages both using machine translation (as presented in [35]), and using multilingual CCA to semantically align representations across languages using bilingual dictionaries from [33]. We compare these two different approaches on three novel extrinsic evaluation tasks, namely, on concept retrieval (Section 5), concept clustering (Section 6) and concept sentiment prediction (Section 7).

Studies on multimodal distributional semantics have combined visual and textual features to learn visually grounded word embeddings and have used the notion of semantics [11, 12] and visual similarity to evaluate them [13,14]. In contrast, our focus is on the visual semantic similarity of concepts across multiple languages which, to our knowledge, has not been considered before. Furthermore, there are studies which have combined language and vision for image caption generation and retrieval [15,16,18,19] based on multimodal neural language models. Our proposed evaluation metric described later in Section 5 can be used for learning or selecting more informed multimodal embeddings which can benefit these systems. Another related study to ours is [20] which aimed to learn visually grounded word embeddings to capture visual notions of semantic relatedness using abstract visual scenes. Here, we focus on learning representations of visual sentiment concepts and we define visual semantic relatedness based on real-world images annotated by community users of Flickr instead of abstract scenes.

## 3 Dataset: Multilingual Visual Sentiment Ontology

We base our study on the MVSO dataset [1], which is the largest dataset of hierarchically organized visual sentiment concepts consisting of adjective-noun pairs (ANPs). MVSO contains 15,600 concepts such as *happy dog* and *beautiful face* from 12 languages, and it is a valuable resource which has been previously used for tasks such as sentiment classification, visual sentiment concept detection, multi-task visual recognition [1,35,40,37]. One shortcoming of MVSO is that the sentiment scores assigned to each affective visual concept was automatically computed through sentiment analysis tools. Although such tools have achieved impressive performances in the recent years, they are typically based on text modalities alone. To counter this, we designed a crowdsourcing experiment with CrowdFlower[1] to annotate the sentiment of the multilingual ANPs in MVSO. We considered 11 out of 12 languages in MVSO, leaving out Persian due to the limited number of ANPs. We constructed separate sentiment annotation tasks for each language, using all ANPs in MVSO for that language.

[1] http://www.crowdflower.com

| | Turkish | Russian | Polish | German | Chinese | Arabic | French | Spanish | Italian | English | Dutch | Average |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Agreement | 66% | 66% | 76% | 69% | 71% | 61% | 65% | 66% | 66% | 70% | 69% | 68% |
| Deviation | 0.77 | 0.70 | 0.39 | 0.59 | 0.54 | 0.79 | 0.67 | 0.59 | 0.58 | 0.48 | 0.52 | 0.60 |

**Table 1** Results of the visual concept sentiment annotations: average percentage agreement and average deviation from the mean score.

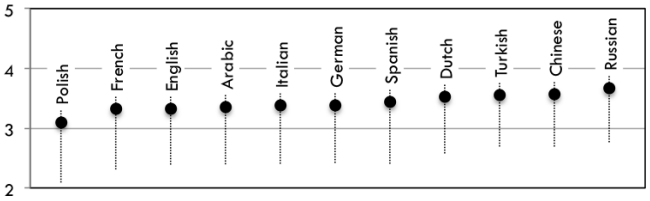


**Fig. 2** Variation of sentiment across languages. The y-axis is the average sentiment of visual concepts in each language (ascending order).

### 3.1 Crowdsourcing Visual Sentiment of Concepts from Different Languages

We asked crowdsourcing workers to evaluate the sentiment value of each ANP on a scale from 1 to 5. We provided annotators with intuitive instructions, along with examples ANPs with different sentiment values. Each task showed five ANPs from a given language along with Flickr images associated with each of those ANPs. Annotators rated the sentiment expressed by each ANP, choosing between "very negative," "slightly negative," "neutral," "slightly positive" or "very positive" with the corresponding sentiment scores ranging from 1 to 5.

The sentiment of each ANP was judged by five or more independent workers. Similar to the MVSO setup, we required that workers were both native speakers of the task's language and highly ranked on the platform.

We also developed a subset of screening questions with an expert-labeled gold standard: to access a crowdsourcing task, workers needed to correctly answer 7 of 10 test questions. To pre-label the sentiment of ANP samples for screening questions, we rank ANPs for each language based on the sentiment value assigned by automatic tools, then use the top 10 ANPs and the bottom 10 for positive/very positive examples and negative/very negative examples respectively. Their performance was also monitored throughout the task by randomly inserting a screening question in each task.

### 3.2 Visual Sentiment Crowdsourcing Results

To assess the quality of the collected annotations of the sentiment scores of ANP concepts, we computed the level of agreement between contributors (Table 1). Although sentiment assessment is intrinsically a subjective task, we found an average agreement around 68% and the agreement percentage is relatively consistent over different languages. We also report results of the mean distance between the average judgement for an ANP and the individual judgements for that ANP: overall, we find that such distance is lower than one, out of a total range of 5.

We found an average correlation of 0.54 between crowdsourced sentiment scores and the automatically assigned sentiment scores in [1]. Although this value is reasonably high, it still shows that the two sets of scores do not completely overlap. A high-level summary of the average sentiment collected per language is shown in Fig. 2. We observe that for all languages there is a tendency towards positive sentiment. This finding is compatible with previous studies showing that there is a universal positivity bias in human language as in [58] and our initial study [1] which was based on automatic sentiment computed from text only, Spanish is found to be the most relatively positive language. **Interestingly, however, here we find that when we combine human language with visual content in the annotation task (as described above), the Russian and Chinese languages carry the most positive sentiment on average when compared to other languages**. This suggests that the visual content has an effect on the degree of positivity expressed in languages.

## 4 Multilingual Visual Concept Matching

To achieve the goal of analyzing the commonality or difference among concepts in different languages, we need a basic tool to represent such visual concepts and to compute similarity or distance among them. In this section, we present two approaches, one based on translation of concepts into a pivot language, and the other based on word embedding trained with unsupervised learning.

### 4.1 Exact Concept Matching

Let assume a set of ANP concepts in multiple languages $\mathcal{C}$ = $\{c_i^{(l)} \mid l = 1 \dots m, i = 1 \dots n_l\}$, where $m$ is the number of languages, $c_i^{(l)}$ is the $i_{th}$ concept out of $n_l$ concepts in the $l_{th}$ language $l$. Each concept $c_i^{(l)}$ is generally a short word phrase ranging from two to five words. To match visual sentiment ANP concepts across languages we first translated them from each language to the concepts of a pivot language using the Google Translate API[2]. We selected English as the pivot language because it has the most complete translation resources (parallel corpora) for each of the other languages due to its popularity in relevant studies. Having translated all concepts to English, we applied lower-casing

[2] `https://cloud.google.com/translate`

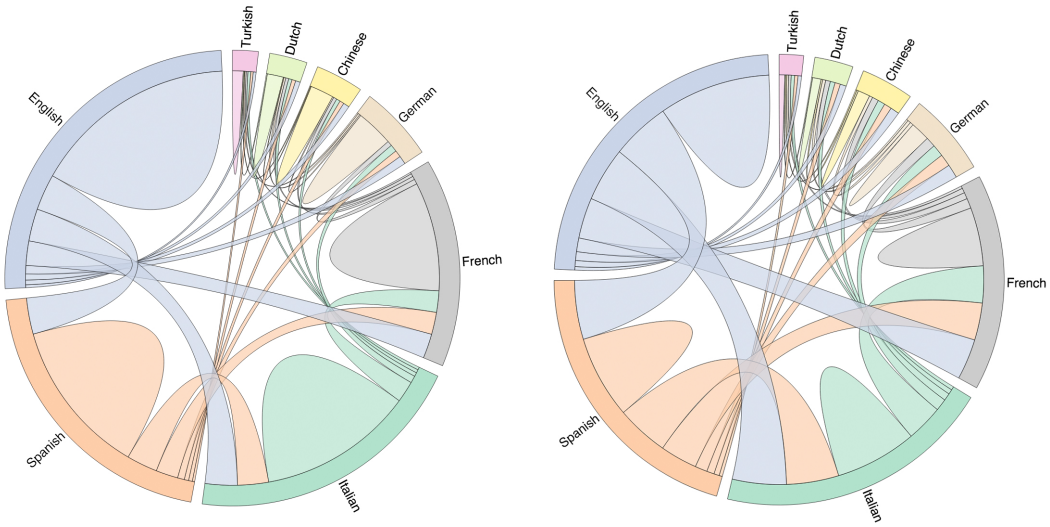


**Fig. 3** Clustering connectivity across top-8 most popular languages in MVSO measured by the number of concepts in the same cluster of a given language with other languages represented in a chord diagram. On the left (a), the clusters based on exact matching are mostly dominated by a single language, while on the right (b), based on approximate matching, connectivity across languages greatly increases and thus allows for more thorough comparison among multilingual concepts.

to all translations and then matched them based on exact-match string comparison.[3] For instance, the concepts *chien heureux* (French), *perro feliz* (Spanish) and *glücklicher hund* (German) are translated to the English concept *happy dog*. Rightly so, one would expect that the visual sentiment concepts in the pivot language might have shifted in terms of sentiment and meaning as a result of the translation process. And so, we examine and analyze the effects of translation to the sentiment and meaning of the multilingual concepts as well as the matching coverage across languages.

### *4.1.1 Sentiment Shift*

To quantitatively examine the effect of translation on the sentiment score of concepts, we used the crowdsourced sentiment values and count the number of concepts for which the sign of the sentiment score shifted after translation in English. We take into account only the translated concepts for which we have crowdsourced sentiment scores; we assume that the rest have not changed sentiment sign. The higher this number for a given language, the higher the specificitiy of the visual sentiment for that language. To avoid counting sentiment shifts caused by small sentiment values, we define a boolean function $f$ based on the crowdsourced sentiment value $\mathrm{s}(\cdot)$ of a concept before translation $c_i$ and after translation $\bar{c}_i$ with a sign shift and a threshold $t$ below which we do not consider sign changes, as follows:

$$f(c_i, \bar{c}_i, t) = |\mathrm{s}(c_i) - \mathrm{s}(\bar{c}_i)| > t. \quad (1)$$

[3] We did not perform lemmatization or any other pre-processing step to preserve the original visual concept properties.

| **Language** | $t = 0.0$ | $t = 0.1$ | $t = 0.2$ | $t = 0.3$ |
|---|---|---|---|---|
| Spanish | 29.1 (6.7) | **16.6** (3.9) | **11.4** (2.6) | **10.1** (2.3) |
| Italian | 28.9 (6.0) | **16.7** (3.3) | **11.4** (2.4) | 7.3 (2.2) |
| French | 36.2 (8.1) | **23.6** (5.3) | **16.8** (3.8) | 9.7 (3.3) |
| Chinese | 24.4 (6.3) | **11.8** (5.5) | 5.5 (1.4) | 3.1 (0.8) |
| German | 27.1 (6.2) | **15.5** (3.5) | 8.3 (1.9) | 7.7 (1.8) |
| Dutch | 18.6 (5.4) | 8.2 (2.4) | 6.2 (1.8) | 3.1 (0.9) |
| Russian | 25.6 (8.3) | **20.5** (6.6) | 5.1 (1.7) | 2.6 (0.8) |
| Turkish | 33.3 (8.2) | **22.2** (5.5) | 7.4 (1.8) | 3.7 (0.9) |
| Polish | 55.5 (16.1) | **38.8** (11.3) | **27.7** (8.1) | **16.6** (4.8) |
| Arabic | 60.0 (21.4) | **40.0** (14.3) | **10.0** (3.6) | **10.0** (3.6) |

**Table 2** Percentage of concepts with sentiment sign shift after translation into English, when using only concepts with crowdsourced sentiment in the calculation or when using all concepts in the calculation (crowdsourced or not). Percentages with significant sentiment shift ($t \geq 0.1$) are marked in **bold**.

For instance when $t > 0$ then all concepts with a sign shift are counted. Similarly, when $t > 0.3$, then only concepts with sentiment greater than 0.3 and lower than -0.3 are counted. These have more significant sentiment sign shift as compared to the ones that fall in to the excluded range.

Table 2 displays the percentage of concepts with shifted sign due to translation. The percentages are on average about 33% for $t = 0$. **The highest percentage of sentiment polarity (sign) shift during translation is 60% from Arabic and the lowest percentage is 18.6% for Dutch.** Moreover, the percentage of concepts with shifted sign decreases for most languages as we increase the absolute sentiment value threshold $t$ from 0 to 0.3. This result is particularly interesting since it suggests that visual sentiment understanding can be enriched by considering the language dimension. We further study this effect on language-specific and crosslingual visual sentiment prediction, in Section 7.

*4.1.2 Meaning Shift and Aligned Concept Embeddings*

The translation can affect also the meaning of the original concept in the pivot language. For instance, a concept in the original language which has intricate compound words (adjective and noun) could be translated to simpler compound words. This might be due to the lack of expressivity of the pivot language, or to compound words with shifted meaning, because of translation mistake, language idioms, or lack of large enough context. For example, 民主法治 (Chinese) is translated to *democracy and the rule of law* in English, while *passo grande* (Italian) is translated to *plunge* and *marode schönheit* (German) is translated in to *ramshackle beauty*.

Examining the extent of this effect intrinsically through, for instance, a cross-lingual similarity task for all concepts is costly because it requires language experts from all languages at hand. Furthermore, the results may not necessarily generalize to extrinsic tasks [21]. However, we can examine the translation effect extrinsically on downstream tasks, for instance by representing each translated concept $c_i$ with a sum of word vectors (adjective and noun) based on $d$-dimensional word embeddings in English, hence $c_i \in R^d$. Our goal is to compare such concept representations which rely on the translation to a pivot language, noted as *translated*, with multilingual word representations based on bilingual dictionaries [33]. In the latter case, each concept in the original language $c_i$ is also represented by a sum of word vectors this time based on $d$-dimensional word embeddings in the original language. These language-specific representations have emerged from monolingual corpora using a skip-gram model (from word2vec toolkit), and have been aligned based on bilingual dictionaries into a single shared embedding space using CCA [17], noted as *aligned*. CCA achieves that by learning transformation matrices $\mathbf{V}, \mathbf{W}$ for a pair of languages which are used to project their word representations $\mathbf{\Sigma}, \mathbf{\Omega}$ to a new space $\mathbf{\Sigma}^*, \mathbf{\Omega}^*$ which can be seen as the shared space. In the multilingual case, every language is projected to a shared space with English ($\mathbf{\Sigma}^*$) space through projection $\mathbf{W}$. The aligned representations have kept the word properties and relation which emerge in a particular language (via monolingual corpora), and at the same time they are comparable with words in other languages (via a shared space). This is not necessarily the case for representations based on translations, because they are trained on a single language.

In Sections 5, 6, 7, we study the translation effect extrinsically on three tasks, namely on concept retrieval, clustering and sentiment prediction respectively. To compare the representations based on translation to a pivot language and representations which are aligned across languages we use the pre-trained aligned embeddings of 512 dimensions based on multiCCA from [33], which were initially trained with a window $w = 5$ on the Leipzig Corpora Collection [34][4].

## 4.2 Matching Coverage

The matching coverage is an essential property for multilingual concept matching and clustering. To examine this property, we first performed a simple clustering of multilingual concepts based on exact matching. In this approach, each cluster is comprised of multilingual concepts which have the same English translation. Next, we count the number of concepts between two languages that belong to the same cluster. This reveals the connectivity of language clusters based on exact matching, as shown in Fig. 3(a) for the top-8 most popular languages in MVSO. From the connection stripes which represent the number of concepts between two languages, we can observe that, when using exact matching, concept clusters are dominated by single languages. For instance, in all the languages there is a connecting stripe that connects back to the same language: this indicates that many clusters contain monolingual concepts. Another disadvantage of exact matching is that out of all the German translations (781), the ones matched with Dutch concepts (39) were more numerous than the ones matched with Chinese concepts (23). This was striking given that there were less (340) translations from Dutch than from Chinese (472). We observed that the matching of concepts among languages is generally very sparse and does not depend necessarily on the number of translated concepts; this hinders our ability to compare concepts across languages in a unified manner. Moreover, we would like to be able to know the relation among concepts from original languages where we cannot have a direct translation.

## 4.3 Approximate Concept Matching

To overcome the limitations of exact concept matching, we relax the exact condition for matching multilingual concepts, and instead we *approximately* match concepts based on their semantic meaning. We performed $k$-means clustering with Euclidean distance on the set of multilingual concepts $\mathcal{C}$ with each concept $i$ in language $l$ being represented by a *translated* concept vector $c_i^{(l)} \in R^d$. Intuitively, in order to match concepts from different languages, we need a proximity (or distance) measure reflecting how 'close' or similar concepts are in the semantic distance space. This enables to achieve our main goal: comparing visual concepts cross-lingually, and cluster them in to multilingual groups. Using this approach, we observed a larger intersection between languages, where German and Dutch share 118 clusters, and German and Chinese intersect over 101 ANP clusters.

[4] http://corpora2.informatik.uni-leipzig.de/download.html

| Language | # Concepts | # Concept Pairs | # Images |
|---|---|---|---|
| English (EN) | 4,421 | 1,109,467 | 447,997 |
| Spanish (ES) | 3,381 | 97,862 | 37,528 |
| Italian (IT) | 3,349 | 44,794 | 25,664 |
| French (FR) | 2,349 | 34,747 | 16,807 |
| Chinese (ZH) | 504 | 21,049 | 5,562 |
| German (DE) | 804 | 14,635 | 7,335 |
| Dutch (NL) | 348 | 3,491 | 2,226 |
| Russian (RU) | 129 | 1,536 | 800 |
| Turkish (TR) | 231 | 941 | 638 |
| Polish (PL) | 63 | 727 | 477 |
| Persian (FA) | 15 | 56 | 34 |
| Arabic (AR) | 29 | 46 | 23 |

**Table 3** ANP co-occurrence statistics for 12 languages, namely the number of concept tags and number of images with concept tags.

When using approximate matching based on word embeddings trained on Google News (300-dimensions), the clustering connectivity between languages is greatly enriched, as shown in Fig. 3 (b): connection stripes are more evenly distributed for all languages. To compute the connectivity, we set the number of clusters $k = 4500$, but we also tried several other values for $k$ which yielded similar results. To learn such representations of meaning we make use of the recent advances in distributional lexical semantics [4, 5, 21, 22] utilizing the skip-gram model provided by word2vec toolkit[5] trained on large text corpora.

#### *4.3.1 Word Embedding Representations*

To represent words in a semantic space we use unsupervised word embeddings based on the skip-gram model via word2vec. Essentially, the skip-gram model aims to learn vector representations for words by predicting the context of a word in a large corpus. The context is defined as a window of $w$ words before and $w$ words after the current word. We consider the following corpora in English on which the skip-gram model is trained:

1. **Google News**: A news corpus which contains 100 billion tokens and 3,000,000 unique words which have at least five occurrences from [43]. News describe real-world events and typically contain proper word usage; however, they often have indirect relevance to visual content.
2. **Wikipedia**: A corpus of Wikipedia articles which contains 1.74 billion tokens and 693,056 unique words which have at least 10 occurrences. The pre-processed text of this corpus was obtained from [24]. Wikipedia articles are more thorough descriptions of real-world events, entities, objects and concepts. Similar to Google News, the visual content is indirectly connected to the word usage.
3. **Wikipedia + Reuters + Wall Street Journal**: A mixture corpus of Wikipedia articles, Wall Street Journal (WSJ) and Reuters news which contains 1.96 billion tokens and 960,494 unique words which have at least 10 occurrences. The pre-processed text of this corpus was obtained from [24]. This combination of news articles and Wikipedia articles captures a balance between these two different types of word usage.
4. **Flickr 100M**: A corpus of image metadata which contains 0.75 billion tokens and 693,056 unique words (with frequency higher than 10) available from Yahoo! [6]. In contrast to the previous corpora, the description of real-world images contains spontaneous word usage which is directly related to visual content. Hence, we expect it to provide embeddings able to capture visual properties.

For the Google News corpus, we used pre-trained embeddings of 300 dimensions with a context window of 5 words provided by [43]. For the other corpora, we trained the skip-gram model with a context window $w$ of 5 and 10 words, fixing the dimensionality of the word embeddings to 300 dimensions. In addition to training the vanilla skip-gram model on word tokens, we also train each of the corpora (except Google News due to lack of access to original documents used for training) by treating each ANP concept as a unique token. This pre-processing step allows the skip-gram model to directly learn ANP concept embeddings while taking advantage from the word contextual information over the above corpora.

#### *4.3.2 Embedding-based Concept Representations*

To represent concepts in a semantic space we use the word embeddings in the pivot language (English) for the *translated* concept vectors, and the aligned word embeddings in the original language for the *aligned* concept vectors. In both cases, we compose the representation of a concept based on its compound words. Each sentiment-biased visual concept $c_i$ comprises zero or more adjective and one or more noun words (as translation does not necessarily preserve the adjective-noun pair structure of the original phrase). Given the word vector embeddings of adjective and noun, $\mathbf{x}_{\text{adj}}$ and $\mathbf{x}_{\text{noun}}$, we compute the concept embedding $\mathbf{c}_i$ using the sum operation for composition ($g$):

$$\mathbf{c}_i = g(x_{\text{adj}}, \mathbf{x}_{\text{noun}}) = \mathbf{x}_{\text{adj}} + \mathbf{x}_{\text{noun}} \quad (2)$$

or the concept embedding $\mathbf{c}_i$ which is directly learned from the skip-gram model. In case of more than two words, say $T$, we use the following formula: $\mathbf{c}_i = \sum_{j=1}^{T} \mathbf{x}_j$. This enables the distance comparison, here with cosine distance metric (see also Section 5), of multilingual concepts using the word embeddings of a pivot language (English) or using aligned word embeddings. At this stage, we note that there are several other ways to define composition of short phrases, e.g. [25,

[5] https://code.google.com/p/word2vec

[6] http://webscope.sandbox.yahoo.com

| Method \ Language | EN | ES | IT | FR | ZH | DE | NL | RU | TR | PL | FA | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| wiki ($w$=10) | 3.81 | 5.62 | 6.47 | 7.18 | 5.30 | 8.33 | 11.65 | 14.67 | 19.59 | 16.62 | 17.25 | 31.17 |
| wiki-anp ($w$=10) | 3.46 | 5.38 | 6.33 | 7.20 | 4.98 | 8.56 | 11.99 | 15.26 | 20.97 | 17.14 | 19.31 | 35.15 |
| wiki-anp-l ($w$=10) | 3.27 | 4.78 | 6.49 | 7.29 | 4.57 | 8.57 | 13.54 | 16.05 | 24.30 | 22.05 | 21.47 | 38.40 |
| wiki_rw ($w$=10) | 10.17 | 12.01 | 12.08 | 12.11 | 13.62 | 12.98 | 11.02 | 13.74 | **12.71** | **12.28** | **6.51** | **16.16** |
| wiki_rw-anp ($w$=10) | 3.79 | 5.54 | 6.38 | 7.23 | 5.16 | 8.53 | 11.67 | 14.94 | 19.79 | 16.48 | 17.91 | 32.34 |
| wiki_rw-anp-l ($w$=10) | 3.57 | 4.90 | 6.43 | 7.21 | 4.90 | 7.91 | 13.28 | 15.27 | 23.29 | 21.15 | 20.15 | 34.59 |
| flickr ($w$=10) | 6.27 | 6.75 | 7.23 | 7.84 | 6.91 | 9.03 | **10.31** | **13.59** | 15.83 | 13.41 | 10.36 | 24.98 |
| flickr-anp ($w$=10) | 3.38 | 4.81 | 6.89 | 6.59 | 4.69 | **7.85** | 11.33 | 14.05 | 18.66 | 16.26 | 15.61 | 31.43 |
| flickr-anp-l ($w$=10) | 2.72 | 4.12 | 5.95 | 6.73 | 4.04 | 8.55 | 14.09 | 14.59 | 25.00 | 22.23 | 21.12 | 34.92 |
| gnews ($w$=5) | 4.59 | 5.81 | 6.85 | 7.51 | 5.63 | 8.76 | 11.08 | 14.02 | 18.29 | 14.88 | 14.08 | 28.61 |
| wiki ($w$=5) | 3.01 | 5.08 | 6.16 | 7.04 | 4.83 | 8.30 | 12.34 | 15.07 | 21.16 | 17.57 | 19.30 | 35.43 |
| wiki-anp ($w$=5) | 2.91 | 5.01 | 6.09 | 7.10 | 4.71 | 8.36 | 12.39 | 15.53 | 21.91 | 17.79 | 20.86 | 37.42 |
| wiki-anp-l ($w$=5) | 2.73 | 4.56 | 6.36 | 7.23 | 4.30 | 8.42 | 13.71 | 16.33 | 25.06 | 22.66 | 22.40 | 40.53 |
| wiki_rw ($w$=5) | 5.70 | 7.36 | 8.12 | 8.47 | 8.51 | 9.48 | 10.52 | 13.60 | 15.34 | 13.43 | 10.12 | 22.40 |
| wiki_rw-anp ($w$=5) | 3.20 | 4.99 | 6.08 | 7.04 | 4.65 | 8.32 | 12.22 | 15.21 | 21.37 | 17.49 | 19.26 | 36.30 |
| wiki_rw-anp-l ($w$=5) | 3.03 | 4.58 | 6.35 | 7.21 | 4.55 | 8.47 | 13.74 | 15.78 | 24.50 | 22.18 | 21.24 | 37.86 |
| flickr ($w$=5) | 5.48 | 6.19 | 6.79 | 7.53 | 6.19 | 8.79 | 10.64 | 13.71 | 16.60 | 14.03 | 11.87 | 28.04 |
| flickr-anp ($w$=5) | 2.87 | 4.52 | **5.85** | **6.56** | 4.41 | 7.91 | 11.85 | 14.34 | 20.18 | 17.14 | 16.67 | 34.31 |
| flickr-anp-l ($w$=5) | **2.21** | **4.12** | 6.04 | 6.84 | **3.94** | 8.28 | 14.66 | 15.54 | 26.10 | 23.16 | 21.82 | 36.85 |

**Table 4** Comparison of the various concept embeddings on visual semantic relatedness per language in terms of MSE (%). The embeddings are from Flickr ('flickr'), Wikipedia ('wiki') and Wikipedia + Reuters + Wall Street Journal ('wiki-rw') trained on a context window of $w \in \{10, 5\}$ words using words as tokens or words and ANPs as tokens ('-anp'). All embeddings use the sum of noun and adjective vectors to compose ANP embedding for a given ANP, except the ones abbreviated with '-anp-l' which use the learned ANP embeddings when available i.e for ANPs which are included in the word2vec vocabulary, and the sum of noun and adjective for those ANPs which are not included in the word2vec vocabulary due to low frequency (less than 100 images). The lowest score per language is marked in bold.

26,43]; however, in this work, we focus on evaluating the type of corpora used for obtaining word embeddings rather than on the composition function.

## 5 Application: Multilingual Visual Concept Retrieval

Evaluating word embeddings learned from text is typically performed on tasks such as semantic relatedness, syntactic relations and analogy relations [4]. These tasks are not able to capture concept properties related to visual content. For instance, while *deserted beach* and *lonely person* seem unrelated according to text, in the context of an image they share visual semantics. An individual person in a deserted beach gives to a remote observer the impression of loneliness. To evaluate various proposed concept representations (namely different embeddings with different training corpora described in Section 4.3.2) on multilingual visual concept retrieval, we propose a ground-truth visual semantic distance, and evaluate which of them retrieves the most similar or related concepts for each of the visual concepts according to this metric.

### 5.1 Visual Semantic Relatedness Distance

To obtain a groundtruth for defining the visual semantic distance between two ANP concepts, we collected co-occurrence statistics of ANP concepts translated in English from 12 languages by analyzing the MVSO image tags (1,000 samples per concept), as shown in Table 3. The co-occurence statistics are computed for each language seperately from each language-specific subset of MVSO. We obtain a visually anchored semantic metric for each language $l$ through the cosine distance between two co-occurrence vectors (k-hot vector containing co-occurence counts) $h_i^{(l)}$ and $h_j^{(l)}$ associated with concepts $c_i^{(l)}$ and $c_j^{(l)}$:

$$d(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)}) = 1 - \frac{\mathbf{h}_i^{(l)} \cdot \mathbf{h}_j^{(l)}}{||\mathbf{h}_i^{(l)}||\,||\mathbf{h}_j^{(l)}||}. \quad (3)$$

The rationale of the above semantic relatedness distance is that if two ANP concepts appear frequently in the same images, they are highly related in the visual semantics and this their distance should be small. We now compare the performance of the various concept embeddings of Section 4.3.1 on the visual semantic relatedness task. Fig. 4 displays their performance over all languages in terms of Mean Squared Error (MSE), and Table 4 displays their performance per language $l$ according to the MSE score for all the pairs of concept embeddings $\mathbf{c}_i^{(l)}$ and $\mathbf{c}_j^{(l)}$, as follows:

$$\frac{1}{T}\sum_{i}^{N}\sum_{j:j\neq i\ \&\ Uij\neq 0}^{|\{i,\ldots,N\}|}\left(d(\mathbf{c}_i^{(l)}, \mathbf{c}_j^{(l)}) - d(\mathbf{h}_i^{(l)}, \mathbf{h}_j^{(l)})\right)^2, \quad (4)$$

where $U_{ij}$ is the co-occurrence between concepts $i$ and $j$, and $T$ is the total number of comparisons, that is:

$$\frac{1}{2}(N^2 - N - |\{Uij = 0\}|). \quad (5)$$

| Method \ Language | EN | ES | IT | FR | ZH | DE | NL | RU | TR | PL | FA | AR |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| Translated concepts ($w$=5) | 5.94 | 4.86 | 5.49 | 5.23 | 5.41 | 6.27 | 7.96 | 13.50 | 11.72 | - | - | - |
| Aligned concepts ($w$=5) | 5.94 | 3.05 | 3.77 | 4.20 | 2.22 | 4.08 | 6.60 | 17.83 | 15.85 | - | - | - |
| Improvement (%) | +0.0 | +59.3 | +45.6 | +24.5 | +143.6 | +53.6 | +20.6 | -32.0 | -35.2 | - | - | - |

**Table 5** Comparison between the translated concepts and the aligned concepts on visual semantic relatedness per language in terms of MSE (%). All embeddings use the sum operation of noun and adjective vectors to compose ANP embedding for a given ANP.

This error function estimates how well the distance defined over the embedded vector concept representation in a given language, $c^{(}l)_i$, can approximate the language-specific visual semantic relatedness distance defined earlier. As seen above, only concept pairs that have non-zero co-occurrence statistics are included in the error function.

## 5.2 Evaluation Results

The highest performance in terms of MSE over all languages (Fig. 4) is achieved by the flickr-anp-l ($w$=5) embeddings, followed by the wiki-anp-l ($w$=5, where $w$ is the window size used in training the embedding) embeddings. The superior performance of flickr-anp-l ($w$=5) is attributed to its ability to learn directly the embedding of a given ANP concept. The lowest performance is observed by wiki-reu-wsj ($w$=10) and flickr ($w$=10). The larger context ($w$=10) performed worse than the smaller context ($w$=5); it appears that the semantic relatedness prediction over all languages does not benefit from large contexts. When the concept embeddings are evaluated per language in Table 4 we obtain slightly different ranking of the methods. In the languages with the most data, namely English (EN), Spanish (ES), Italian (IT), French (FR) and Chinese (ZH), the ranking is similar as before, with flickr-anp-l ($w$=5), flickr-anp ($w$=5) and wiki-anp ($w$=5), wiki-anp-l ($w$=5) embeddings having the lowest error in predicting semantic relatedness.

Generally, we observed that for well-resourced languages the quality of concept embeddings learned by a skip-gram model improves when the model is trained using ANPs as tokens (both when using directly learned concept embeddings or composition of word embeddings with sum operation). Furthermore, the usage of learned embeddings abbreviated with $-l$ on the top-5 languages outperforms on average all other embeddings in English, Spanish and Chinese languages and performs similar to the best embeddings on Italian and French. In the low resourced languages the results are the following: in German (DE) language the lowest error is from flickr-anp ($w$=10), in the Dutch (NL) and Russian (RU) is the flickr ($w$=10). Lastly, the lowest error in the Turkish (TR), Persian (FA) and Arabic (AR) languages is from wiki-reu-wsj ($w$=10). It appears that for the languages with small data the large context benefits the visual semantic relatedness task.

Moreover, the performance of embeddings with a small context window ($w = 5$), is outperformed by the ones that use a larger one ($w = 10$) as the number of image examples of the languages decreases. This is likely due to the different properties which are captured by different context windows, namely more abstract semantic and syntactic relations with a larger context window and more specific relations with a smaller one. Note that the co-occurrence of concepts in MVSO images is computed on the English translations and hence some of the syntactic properties and specific meaning of words of low-resourced languages might have vanished due to errors in the translation process. Lastly, the superior performance of the embeddings learned from the Flickr 100M corpus in the top-5 most resourced languages, validates our hypothesis that word usage directly related to the visual content helps (like the usage in Flickr) learn concept embeddings with visual semantic properties.

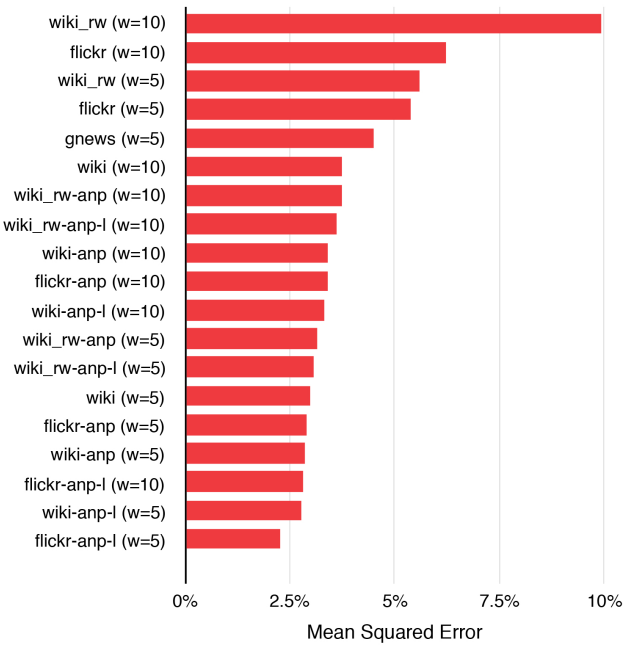


**Fig. 4** Comparison of the various concept embeddings over all languages on visual semantic relatedness in terms of descending MSE (%). For the naming conventions please refer to Table 4.

## 5.3 Translated vs. Aligned Concept Representations

To study the effect of concept translation, we compare on the visual semantic relatedness task the performance of 500-dimensional translated and aligned concept representations both trained with word2vec with a window $w = 5$ on Leip-

sig Corpus (see Section 4.1.2). The evaluation is computed for all the languages which have more than 20 concept pairs the concepts of which belong to the vocabulary of the Leipzig corpus (e.g. PL, AR and FA had less than 5). The results are displayed on Table 5. Overall, the aligned concept representations perform better than the translated ones on the languages with a high number of concept pairs (more than 40), namely, Spanish, Italian, French, Chinese, German and Dutch, while for the low-resourced languages, namely, Russian and Turkish, they are outperformed by the translated concept representations. The greatest improvement of aligned versus translated representations is observed on the Chinese language (+143%), followed by Spanish (+59%), German (+53%) and Italian (+45%), and the lowest improvement is on French (+24%) and Dutch (+20%). These results show that the translated concepts to English do not capture all the desired language-specific semantic properties of concepts, likely because of the small-context translation and the English-oriented training of word embeddings. Furthermore, the results suggest that the concept retrieval performance of all the methods compared in the previous section will most likely benefit from a multilingual semantic alignment. In the upcoming sections, we will still use the translated vectors to provide a thorough comparison across different training tasks and further support the above finding.

## 6 Application: Multilingual Visual Concept Clustering

Given a common way to represent multilingual concepts, we are now able to cluster them. As discussed in Section 4, clustering multilingual concept vectors makes it easier to surface commonly shared concepts (when all languages present in a cluster) versus concepts that persistently stay mono-lingual. We experimented with two types of clustering approaches: a *one-stage* and a *two-stage* approach. We also created a user interface for the whole multilingual corpora of thousands of concepts and images associated with them based on the results of these clustering experiments [1]. This ontology browser aligns the images associated with semantically close concepts from different cultures.

### 6.1 Clustering Methods

The *one-stage* approach directly clusters all the concept vectors using $k$-means. The *two-stage* clustering operates first on the noun or adjective word vectors and then on concept vectors. For the two-stage clustering, we perform part-of-speech tagging on the translation to extract the representative noun or adjective with TreeTagger [27]. Here, we first cluster the translated concepts based on their noun vectors only, and then run another round of $k$-means clustering within the clusters formed in the first stage using the vector for the full concept. In the case when a translation phrase has more than one noun, we select the last noun as the representative and use it in the first stage of clustering. The second stage uses the sum of vectors for all the words in the concept. We also experimented with first clustering based on adjectives and then by full embedding vector using the same process. In all methods, we normalize the concept vectors to perform $k$-means clustering over Euclidean distances.

| Method | Embeddings | $sen_C$ | $sem_C$ | $\mu$ |
|---|---|---|---|---|
| 2-stage_noun | gnews (w=5) | 0.278 | 0.676 | 0.477 |
| 2-stage_adj | gnews (w=5) | **0.161** | 0.614 | **0.388** |
| 1-stage | wiki-anp (w=10) | 0.239 | 0.659 | 0.449 |
| 1-stage | wiki_rw-anp (w=10) | 0.242 | 0.582 | 0.412 |
| 1-stage | flickr-anp (w=10) | 0.242 | 0.535 | 0.388 |
| 1-stage | wiki-anp (w=5) | 0.239 | 0.659 | 0.449 |
| 1-stage | wiki_rw-anp (w=5) | 0.234 | 0.579 | 0.407 |
| 1-stage | flickr-anp (w=5) | 0.246 | **0.532** | 0.389 |

**Table 6** Sentiment and semantic consistency of the clusters using multilingual embeddings $k$-means clustering methods with $k = 4500$, trained with the various concept embeddings. The full MVSO corpus is used for clustering ( 16K concepts).

We adjust the $k$ parameter in the last stage of two-stage clustering based on the number of concepts enclosed in each first-stage cluster, e.g. concepts in each noun-cluster ranged from 3 to 253 in one setup. This adjustment allowed us to control the total number of clusters formed at the end of two-stage clustering to a target number. With two-stage clustering, we ended up with clusters such as *beautiful music*, *beautiful concert*, *beautiful singer* that maps to concepts like *musique magnifique* (French), *bella musica* or *bellissimo concerto* (Italian). While noun-first clustering brings concepts that talk about similar objects, e.g. estate, unit, property, building, adjective-based clustering yields concepts about similar and closely related emotions, e.g. grateful, festive, joyous, floral, glowing, delightful (these examples are from two-stage clustering with the Google News corpus).

We experimented with the full MVSO dataset (Table 6) and a subset of it which contains only face images (Table 7). From the 11,832 concepts contained in the full MVSO dataset, only 2,345 concepts contained images with faces. To evaluate the clustering of affective visual concepts, we consider two dimensions: (1) *Semantics*: ANPs are concepts, so we seek a clustering method to group ANPs with similar semantic meaning, such as for example *beautiful woman* and *beautiful lady*, (2) *Sentiment*: Given that ANPs have an *affective* bias, we need a clustering method that groups ANPs with similar sentiment values, thus ensuring the integrity of ANPs' sentiment information after clustering.

### 6.2 Evaluation Metrics

To evaluate the clustering of affective visual concepts, we consider two dimensions: (1) *Semantics*: ANPs are concepts,

| Method | Embeddings | $sen_C$ | $sem_C$ | $\mu$ |
|---|---|---|---|---|
| 2-stage_noun | wiki (w=10) | 0.511 | 0.588 | 0.549 |
| 2-stage_noun | wiki_rw (w=10) | 0.529 | 0.604 | 0.566 |
| 2-stage_noun | flickr (w=10) | 0.538 | 0.528 | 0.533 |
| 2-stage_noun | wiki (w=5) | 0.534 | 0.586 | 0.560 |
| 2-stage_noun | wiki_rw (w=5) | 0.510 | 0.614 | 0.562 |
| 2-stage_noun | flickr (w=5) | 0.526 | 0.513 | 0.519 |
| 2-stage_noun | gnews (w=5) | 0.309 | 0.569 | 0.439 |
| 2-stage_adj | wiki (w=10) | 0.483 | 0.567 | 0.524 |
| 2-stage_adj | wiki_rw (w=10) | 0.476 | 0.536 | 0.506 |
| 2-stage_adj | flickr (w=10) | 0.459 | 0.536 | 0.497 |
| 2-stage_adj | wiki (w=5) | 0.581 | 0.930 | 0.755 |
| 2-stage_adj | wiki_rw (w=5) | 0.472 | 0.560 | 0.516 |
| 2-stage_adj | flickr (w=5) | 0.455 | 0.519 | 0.487 |
| 2-stage_adj | gnews (w=5) | **0.178** | 0.522 | **0.350** |
| 1-stage | wiki-anp (w=10) | 0.240 | 0.576 | 0.408 |
| 1-stage | wiki_rw-anp (w=10) | 0.257 | 0.508 | 0.382 |
| 1-stage | flickr-anp (w=10) | 0.262 | **0.489** | 0.375 |
| 1-stage | wiki-anp (w=5) | 0.250 | 0.583 | 0.416 |
| 1-stage | wiki_rw-anp (w=5) | 0.281 | 0.522 | 0.402 |
| 1-stage | flickr-anp (w=5) | 0.280 | 0.502 | 0.391 |

**Table 7** Sentiment and semantic consistency of the clusters using multilingual embeddings $k$-means clustering methods with $k = 1000$, trained with the various concept embeddings. The subset of concepts in *portraits* corpus is used for clustering ( 2.3K concepts).

so we seek a clustering method to group ANPs with similar semantic meaning, such as for example *beautiful woman* and *beautiful lady*, (2) *Sentiment*: Given that ANPs have an *affective* bias, we need a clustering method that groups ANPs with similar sentiment values, thus ensuring the integrity of ANPs sentiment information after clustering.

### *6.2.1 Semantic Consistency*

Each clustering method produces $k$ ANPs clusters, out of which $C$ contains two or more ANPs. For each of these multi-ANP clusters, each with $N_m$ ANPs with $\text{ANP}_{m,j}$ being the $j_{th}$ concept in the $m_{th}$ cluster, we compute the average visually grounded semantic distance (Eq. 4) between all pairs of ANPs, and then we average them over all $C$ clusters, thus obtaining a Semantic Consistency $\text{sem}_C$ metric for a given clustering method:

$$\frac{1}{C}\sum_{m=1}^{C}\frac{1}{N_m}\sum_{j:j\neq i \ \& \ Uij\neq 0}^{|\{i,\ldots,N_m\}|} d(\text{ANP}_{m,i},\text{ANP}_{m,j}), \qquad (6)$$

### *6.2.2 Sentiment Consistency*

. For each multi-ANP cluster $m$, we compute a *sentiment quantization error*, namely the average difference between the sentiment of each ANP in the cluster, and the average sentiment of the cluster. Therefore, given the average sentiment for a cluster $m$, $\text{sen}_m = \sum_{i=1}^{N_m}\text{sen}(\text{ANP}_i)/N_m$ with $\text{ANP}_{m,j}$ being the $j_{th}$ concept in the $m_{th}$ cluster, we obtain

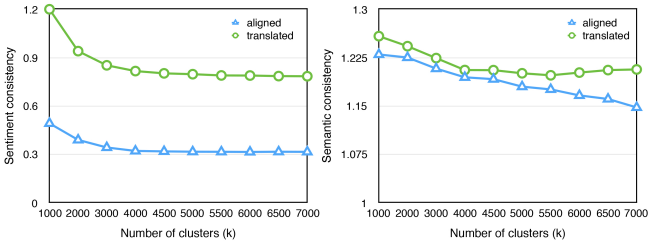


**Fig. 5** Comparison of the aligned (15630 concepts) versus the translated concept embeddings (11834) over all languages on clustering in terms of sentiment (left) and semantic (right) consistencies.

a sentiment consistency metric, noted $sen_C$, for a given clustering method as follows:

$$\frac{1}{C}\sum_{m=1}^{C}\frac{1}{N_m}\sum_{i=1}^{N_m}(\text{sen}(\text{ANP}_{m,i}) - \text{sen}_m)^2 \qquad (7)$$

## 6.3 Evaluation Results

We evaluate all the clustering methods using these two scoring methods and an overall *consistency* metric which is the average of semantic and sentiment consistencies. The lower the value of the metrics, the higher the quality of the clustering method. We observe that semantic consistency and sentiment consistency are actually highly related. When we correlate the vector containing semantic consistency scores for all clustering methods with the vector containing sentiment consistency scores, we find that the Pearson's coefficient is around 0.7, suggesting that the higher the semantic relatedness of the clusters resulting from one method, the higher their respective sentiment coherence. Given that when the number of clusters $k$ increases the average consistency within a cluster generally increases (regardless of the language and training corpus of the embeddings), we avoided very large values for $k$. Based on the results, among the two-stage methods, the adjective-first clustering which uses the Google News embeddings produced the lowest average consistency error. This confirms our intuition that similar sentiments are clustered together when we first cluster ANPs with similar adjectives. Among the one-stage methods, the embeddings trained on Flickr were superior to other corpora. More generally, the embeddings which were trained on full ANP tokens lead to increased semantic consistency, similar to the results presented in Section 5.1.

Results of clustering based on the multilingual aligned concepts [33] and translated concepts, as described in Section 4.1.2 are provided in Fig. 5. We performed language-specific consistency computation in order to compare clustering based on these two methods. In this evaluation method, we first compute the consistency within concepts coming from the same language and then compute the consistency per cluster by averaging language-specific consistencies within

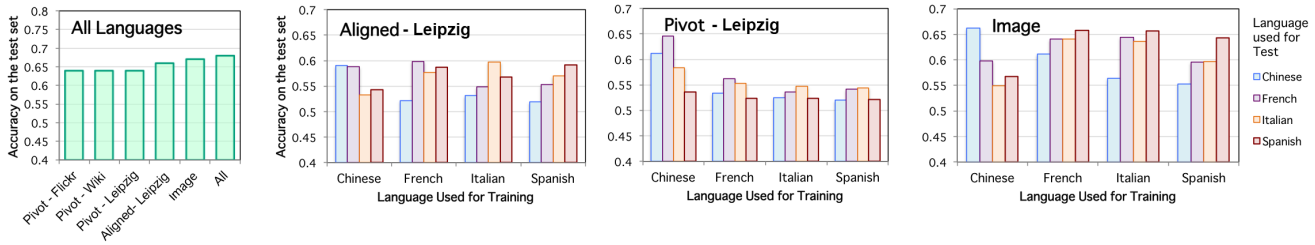


**Fig. 6** Balanced accuracy on corresponding test sets for sentiment prediction over all languages (first plot) and cross-language (other three plots) using visual concept representations (*Image*), textual concept representations (*Aligned*, *Pivot*) using three different domains (Flickr, Wikipedia and Leipzig corpora) and a multimodal combination comprised of *Image* representation and *Aligned - Leipzig* representation (*All*).

that cluster. This is needed in order to be able to compare the clustering based on two different sample sizes of concepts being clustered as well as being able to compare two language-wise different concept embedding spaces: we have 15,630 original ANPs represented in multilingual aligned concepts space instead of 11,834 multilingual translated concepts. We observed that multilingual aligned concepts performed better in both evaluation tasks and did generate more sentimentally consistent clusters.

## 7 Application: Multilingual Visual Sentiment Prediction

To further test the quality of our concept representations, we perform a small prediction experiment. The task is ANP sentiment prediction. Given a concept expressed in the form of adjective-noun pair, we want to build a framework able to automatically score its sentiment. We use and compare various ANP representations: translated concept vectors, aligned concept vectors and visual features. We use representations as features, and crowdsourced sentiment values as annotations, and train a learning algorithm to distinguish between positive and negative visual sentiment values.

### 7.1 Experimental Setup

We consider all ∼15K ANPs annotated with crowdsourced sentiment scores. To reduce sentiment classification to a binary problem (similar to previous work [1]), we discretise continuous sentiment annotations by considering as positive all ANPs whose sentiment is higher than the median sentiment of the whole dataset, and the rest as negative. We then describe their content using three different methods:

- *Translated Concept Vectors:* After translating all concepts to the English language, we compute the sum of the adjective and noun 512-dimensional word embeddings trained on the English version of the Leipzig corpus from [33].
- *Aligned Concept Vectors:* We compute the sum of adjective and noun 512-dimensional aligned word embeddings based on multiCCA from [33] trained on Leipzig corpus.
- *Average Visual Features:* For all images tagged with a given ANP, we extract the 4096 features at the second-to-last layer (fc7) of the CNN designed for visual ANP detection [1]. To get a compact representation of ANPs, we average fc7 features across all images of the ANP.

We then train a random forest classifier with 20, 50 and 100 trees respectively for translated vectors, aligned vectors and visual features (parameters were tuned with cross-validation on the training set), resulting in four trained models for sentiment detection. We evaluate the performances of these models with balanced average accuracy on the test set. We perform two experiments: all-language sentiment classification and cross-language sentiment classification.

In the former, to understand how predictive different kinds of features are for ANPs in any language, we mix ANPs from thelanguages into the same pool and split it into 50% for training an 50% for testing. . We use the translated features (into English) as features and the random forest model as the predictor model. In addition, we also add a model based on combination *Textual and Visual Features* by concatenating the aligned concept vector and the ANP visual features into a single compact feature vector describing ANPs in a multimodal fashion. In the latter, to understand the extent to which features are predictive for sentiment of different languages, we design language-specific sentiment classifiers. We consider four main languages, Chinese, Italian, French and Spanish, and split the corresponding language-specific ANPs into 50-50 train-test. For each language, we then train a separate model, using a random forest classifier with same parameters as above. Similar to previous work [1], we also perform *cross-language* prediction where we use a predictor trained on one language to detect sentiment for ANPs in other languages. This helps us understand not only similarities and differences between different cultures when expressing visual emotions, but also how different modalities (textual and visual) impact such differences.

### 7.2 All language sentiment classification.

From the first subplot of Fig. 6 we can see that, although very different in nature, the examined ANP representations

achieve similar level of accuracy (64% to 68%) on the test set for ANP sentiment prediction, with the visual representation being the most effective and the pivot-based concept representation the least effective. Moreover, we can observe that the aligned representations (align) outperform the ones based on translation (pivot). We also tried to improve translation-based vectors (pivot) using domain-specific corpora such as Flickr which worked well on the retrieval task, however we did not observe significant differences to vectors from the Leipzig corpus. Hence, we concluded that domain-specificity is not that important for this task. Lastly, the accuracy improves when combining different representations into a single multimodal vector.

### 7.3 Cross-language sentiment classification.

We report in the last 3 subplots of Fig. 6 the cross-language sentiment classification results: for each language, we report average accuracy performances for sentiment detection for predictors trained on every separate language. We can see similar patterns to the all-language sentiment classification task: **Visual features are the most predictive for ANP sentiment followed closely by text-based features based on aligned concept representations, while the text-based features based on pivot language are the least predictive.**

Moreover, by looking at visual-based cross-lingual sentiment classification, we can notice the same patterns exposed by previous work on cross-cultural sentiment detection: **It is very difficult to predict the sentiment of ANP concepts of Eastern languages such as Chinese from detectors trained on Western languages. In contrast, cross-lingual sentiment prediction between Latin-based languages (French, Italian, Spanish) tend to perform comparatively well.** While similar patterns can be found for predictions from using aligned vectors, when using vectors generated after translating all ANPs to English, cross-lingual sentiment differences tend to disappear (all the pivot language-specific detectors show the same performances).

## 8 Application: Discovering Interesting Clusters

One application of the proposed framework is to provide a data-driven way for discovering clusters with interesting properties in terms of language coverage, semantic and sentiment consistencies. These "interesting" clusters themselves can shed lights on the inner connections as well as cultural uniqueness across different languages.

In order to evaluate semantic consistency of clusters from different perspectives, we define separate similarity metrics for adjective-level, noun-level and visual comparisons.

### 8.1 Semantic Consistency

Given a cluster $c$ populated with $N_c$ ANPs, we consider the exact English translation of each ANP it contains. We define the intra-cluster semantic similarity of adjectives as follows:

$$sim_{ADJ}(c) = \frac{1}{N_c} \sum_{1 \leq i < j \leq |N_c|} d(ADJ_i, ADJ_j), \tag{8}$$

where $ADJ\_i$ is the word embedding vector for the translated adjective. Similarly, the in-cluster semantic similarity of nouns is defined as $sim_{NOUN}(Cc)$. Lastly, the semantic consistency score of the cluster $C$ is defined as:

$$sim(c) = \alpha \cdot sim_{ADJ}(c) + (1 - \alpha) \cdot sim_{NOUN}(c) \tag{9}$$

where $T$ is the total number of comparisons, and $\alpha$ controls the weights of similarities of adjectives and nouns. In our experiment, $\alpha$ is set as $0.5$.

### 8.2 Visual Consistency

To systematically discover visually similar and dissimilar clusters, we also compute a pure *visual consistency metric*. For each $ANP_{c,i}$ in a cluster $c$, we compute the corresponding Average Visual Feature $AVF_{c,i}$ as shown in Sec. 7.1. We then calculate for a given cluster a visual consistency metric as the average distance between pairs of ANPs in a cluster based on their corresponding visual feature vector.

### 8.3 Multilingual Clusters

To discover interesting visual concepts that different languages tend to talk about when attaching emotion to pictures, we proceed as follows. We rank highly multilingual clusters (including more than 4 languages) according to their semantic consistency, and consider the top 50. This ensures that we are looking at good clusters with highly semantically-related ANPs. By then looking at the sentiment consistency, we can distinguish clusters of concepts for which different languages agree on the sentiment value, versus clusters of concepts with high sentiment diversity, as well as visually similar and dissimilar clusters.

#### *8.3.1 Sentiment-Invariant and Sentiment-Diverse Clusters*

The clusters of concepts the sentiment of which do not vary across language represent somehow those topics for which sentiment value is *universal*. These include highly-positive clusters such as *funny baby, amazing baby, cute baby*, *happy baby*, and *warm hearts*: all languages agree on the positive emotion conveyed by such fundamental concepts.

$$vis(c) = \frac{1}{N_c} \sum_{1 \leq i < j \leq |N_c|} d(AVF_{c,i}, AVF_{c,j}) \tag{10}$$


**Fig. 7** Clusters with diverse sentiments of *historical_building* from Russian, French, English and Dutch.

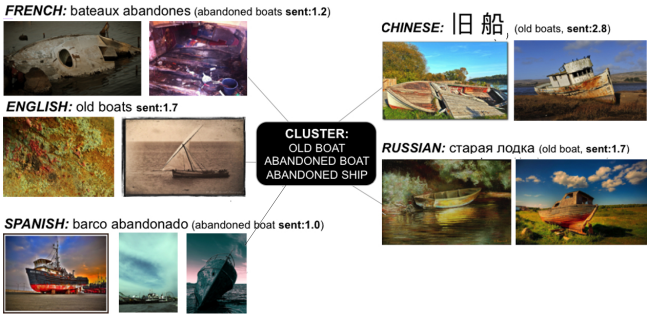


**Fig. 8** Clusters with high sentiment variance between Eastern and Western cultures: *abandoned boats*.

To look at the most *encultured* visual concepts, we compute, for each cluster, the sentiment diversity across languages. We find that the cluster of *historical building* has higher sentiment diversity across cultures, as shown in Fig. 7. A visual inspection suggests that there is a culture-specific image style composition related to this difference in the data: historical buildings are depicted in a more dense and contrasting way in Russian while relatively mild and soft in French, Dutch and English.

### *8.3.2 Eastern vs. Western Sentiment Clusters.*

Another interesting analysis is to observe how sentiment varies across cultures on a given cluster. To study this property, we compute, for each cluster, the difference between the sentiment assigned by the Chinese culture and the average sentiment of Western languages such as Italian, French, Spanish and English. **Interestingly, we found that clusters such as *abandoned boats* or *abandoned house* show huge sentiment variation between Eastern and Western cultures** (see Fig. 10 for a visual example). By looking at the pictures corresponding to the ANPs in each clusters, we understand the reason behind this separation: while Western cultures represent abandoned objects as houses or ships in ruins, Eastern cultures ascribe a romantic sense to the notion of abandonment.


**Fig. 9** Monolingual clusters for five languages, obtained from the 1-stage $k$-means clustering which uses Flickr concept embeddings. The clusters are based on distinctive concepts which reveal cultural insights due to their uniqueness, expressivity and cultural specificity.

## 8.4 Language-Specific Clusters

Apart from the multilingual clusters, it is also important to identify clusters which are based on a single language, because they may reveal culture-specific insights. To achieve this, we simply count the number of languages for each cluster estimated by any clustering method above and filter out all the clusters which contain more than one language. Then we can, for instance, select from the remaining clusters the top-$k$ ones which contain the greatest number of concepts and (or) images and (or) other additional criteria.

However, the selection depends heavily on the value of $k$ in $k$-means, and, therefore, it is hard to find a general rule to discover monolingual clusters unique to a specific language. One way to achieve more reliable clusters is by running a clustering method for several values of $k$, and observe which set of concepts/clusters tend to remain monolingual more often than not. In case that, a cluster fails to accomodate other languages through several values of $k$, it most likely means that it contains culture specific concepts. Fig. 9 displays manually selected examples of monolingual clusters for Spanish, Italian, French and Chinese, obtained from the 1-stage clustering method ($k = 1000$) which uses the Flickr ($w = 10$) embeddings. During our manual inspection, we observed that for many clusters it is quite difficult to decide whether it is specific to a given culture or not, and that few clusters are more distinctive to specific culture according to our judgement after discussion with natives.

### *8.4.1 Visually Similar and Dissimilar Clusters*

To discover the most visually consistent clusters across languages, we select the clusters with low visual consistency. **The clusters of visually similar concepts tend to refer to either historic/religious contexts (e.g. *medieval festival* or *holy communion*), universally positive concepts such as *happy birthday* or *happy people*, or universally negative concepts such as *sexual violence***.

To look at the most visually inconsistent concepts, we select the clusters with high $vis(c)$. We find that visually dissimilar clusters correspond to very subjective notions such


**Fig. 10** Images of the same sentiment concept (e.g. *good food*) show distinct patterns in different languages, here Chinese and Italian. Patterns are discovered automatically using the technique in [60].

as aesthetic appeal, for example *beautiful woman*, *beautiful girl*, *beautiful flowers*. **Visually dissimilar clusters also tend to gather concepts that are very culture specific, such as *healthy food*, *good food* or *famous actress***.

By using visual pattern mining techniques developed in [60], we can further discover distinct visual patterns associated with each language/culture. For example, Figure **??** shows the unique visual patterns associated with the ANP concept "good food" in Chinese and Italian. It's very interesting to observe the culture-specific food or scene patterns, such as wine, pasta for Italian, and mixed fried dish and people for Chinese.

## 9 Case Study: Portrait Concept Clustering

The proposed multilingual concept clustering framework can be a useful tool for exploring and analyzing any large, multilingual collections of visual concepts. As an example application, we applied this framework to study how affective concepts attach to human *portraits*, i.e. photos with faces, through the viewing lens of different languages.

### 9.1 Portrait-based Sentiment Ontology

Portrait and face-centric photography has been a subject of research in multiple disciplines for years. Facial perception is among the most developed human capabilities, where our brains even contain a dedicated sub-network of neurons for face processing [28]. Recently, computational understanding portrait modeling has attracted much attention from the multimedia community, e.g. in computational aesthetics [29], animated GIFs [30], and social dynamics [31]. Here, we seek to unpack what sentiment-biased visual concepts, specifically ANPs, languages attach to faces.

#### *9.1.1 Face Detection and ANP Filtering*

To obtain a corpus of visual concepts relating to faces, we ran a frontal face detector [32] which projects images onto a normalized pixel difference feature space and performs quadtree-based face detection. A total of 3,858,869 faces were detected across the 7,368,364 images in the MVSO image dataset [1]. Over 53.67% of these detections came from the English image subset (2,071,078 detections), where the next leading language subset was Spanish at 23.68% (913,596 detections). We then computed a *portrait score* for each ANP which we define as the ratio of detected faces to all images in each ANP. We then selected the subset of ANPs whose portrait score was greater than 0.6. To ensure statistical significance, we only considered languages with 20 or more face-dominated ANPs: Turkish, Russian, German, Chinese, French, Spanish, Italian and English. Of the 11,832 concepts from the full MVSO dataset, we retained 2,345 face ANPs. We found that in general, detected faces from the French and German datasets are larger in size on average than those of other languages. In addition, images originating from the Italian subset typically contained more than one person while images in the Chinese and Turkish subset tended to contain mostly single-subject portraits.

#### *9.1.2 Concept Sentiment and Face Images*

To explore the sentiment correlations of different languages to the presence of faces, we computed the Pearson's correlation coefficient $\rho$ for each language between ANP portrait scores and the ANP sentiment values, as shown in Table 8. The higher the correlation, the higher the tendency of a given language to associate positive sentiment with a face image. Here, for all languages except Turkish, the presence of portraits in an ANP image pool tended to be positively correlated with the ANP sentiment. In particular, the languages having the strongest tendency to attach positive sentiments to portraits are Russian and Chinese.

| | $\rho$(face,sent) | sent(faces) | sent(all) | diff(%) | face size(%) | #faces(%) |
|---|---|---|---|---|---|---|
| Turkish | 0.00 | 3.54 | 3.55 | -0.26 | 63.46 | 0.95 |
| Russian | **0.23** | **4.13** | 3.67 | **12.48** | 58.25 | 1.23 |
| German | 0.18 | 3.75 | 3.39 | 10.70 | 65.49 | 0.99 |
| Chinese | **0.23** | **4.30** | 3.57 | **20.33** | 64.12 | 0.93 |
| French | 0.14 | 3.48 | 3.32 | 4.79 | **65.88** | 1.01 |
| Spanish | 0.16 | 3.72 | 3.44 | 7.93 | 65.08 | 1.23 |
| Italian | 0.15 | 3.75 | 3.38 | 10.97 | 61.72 | **1.37** |
| English | 0.15 | 3.51 | 3.32 | 5.49 | 56.78 | 1.04 |

**Table 8** Sentiment statistics per language (face and all ANPs).

## 9.2 Multilingual Portraits

We sought to investigate how similar/different languages are with regard to affective visual concepts (ANPs) and their face images. We clustered face ANPs from the subset of eight languages using our approximate match-based clustering techniques, and evaluated different clustering approaches to find that the *single-step* clustering over Flickr-trained vector with $w = 5$ and $k = 1000$ gave us the best results when combining semantic and sentiment consistency metrics. This method output 1,000 multilingual clusters of affective visual concepts related to portraits. This provides a powerful tool to analyze the visual concept preferences for different languages, i.e. if the ANPs of two languages fall often in similar clusters, such languages tend to attach similar concepts to face images.

### *9.2.1 Languages, Sentiments and Face Sizes*

Among the discovered clusters around 60% are monolingual clusters: English and Spanish are the languages with the highest percentage of monolingual clusters (∼32% and 31% respectively), probably due to the large number of ANPs, implying a wider vocabulary than other languages. Around 22% of the face ANP clusters are bilingual, out of which 40% contains 2 of the 4 Western-most languages in the corpus (French, Italian, English and Spanish), while the others contain a mixture of other languages. The remaining 18% of clusters show three languages or more. But what is the relation between multilingual clusters and portrait sentiment? Do languages agree more on positive or negative sentiment for similar visual concepts? To answer these, we computed the correlation between the number of languages falling into each cluster and the average sentiment of the ANPs in that cluster. These two dimensions statistically significantly correlate with a coefficient of 0.13, showing the following: **Different languages tend to associate similar visual concepts to portraits when concepts carry positive sentiment.**

We also analyse the relation between cluster multilingualness and face sizes. Do visual concepts shared by different languages refer to portraits with bigger or smaller faces? We compute here the correlation between the number of languages falling into each cluster and the average face sizes for the ANPs in that cluster. The correlation coefficient stands at 0.17, showing that, the bigger the average face sizes of portraits related to a visual concept, the higher the possibility that different languages share such concept.

### *9.2.2 What Do Visual Sentiment Concepts in Different Languages Reveal About Portraits?*

According to our clustering method, ∼3% of the clusters contain five or more languages. The limited size of this data allows us to proceed with manual inspection, to understand the topics of the ANPs falling in highly multilingual clusters. The most highly multilingual cluster contains 8 languages and 20 ANPs: its main topic is about *little guy* or *little girl* (e.g. *piccola bimba* in Italian, or *petit fille* in French). One interesting observation is that not all languages agree on the sentiment value for this concept: while Chinese and Turkish give a score slightly below 3, Italian, Spanish and Russian languages consider it a very positive concept, having respectively an average sentiment value of 4.0, 4.0 and 4.6, respectively, for this cluster. The second biggest cluster, spanning seven languages (all apart from Chinese) with 24 ANPs, contained concepts like *gorgeous girl* in English or *belle fille* in French. Here, all languages agree on the highly positive sentiment value of this concept . However, as seen in the previous Section, the visual representations of such concepts tend to vary across languages. Other highly multilingual noun clusters contain concepts related to *happy children*, *young women*, *healthy food*, *beautiful women*, etc, and also negative concepts such as *sexual violence*.

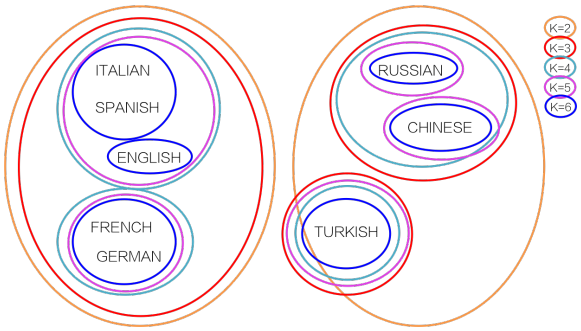


**Fig. 11** Groups of languages according to their similarity when associating affective concepts to faces. Each color corresponds to the output of a different clustering granularity $k$.

## 9.3 FaceMVSO-based Language Clustering

Which languages are more similar when tagging portraits? To further understand language-specific concepts used when tagging face images, we perform a multivariate analysis of language distribution across visual concepts. To better understand which groups of languages tend to attach similar affective concepts to face images, we proceed as follows. We create eight $k$-dimensional vectors, one for each language. Each element of such vectors corresponds to the number of ANPs falling into each cluster, normalized by the total number of ANPs for a given language. Finally, we cluster these vectors using $k$-means with cosine distance, progressively raising $k$ from 2 to 6, thus separating languages into different groups, as shown in Fig. 11. The binary subdivision of the languages in two clusters ($k = 2$) shows immediately a clear separation between more Eastern (Turkish, Chinese, Rus-

sian) and Western (Italian, Spanish, French, English, German) languages. When $k = 3$, Turkish is the first one to detach from the Eastern clusters, suggesting that Turkish images tend to have more unique ways to assign concepts to portraits. Within the Western languages, when $k = 4$ we see a separation: Italian gets clustered with Spanish and English, while French gets clustered with German. Chinese and Russian become independent clusters for $k = 5$. Finally, when $k = 6$, Western languages split again: English becomes an independent cluster, leaving two bilingual clusters: Italian–Spanish and French–German.

## 10 Discussion and Conclusion

In this paper, we developed tools which allow researchers and practitioners to explore the impact of culture in visual sentiment perception. In particular, we proposed a multimodal framework for multilingual visual sentiment concept retrieval and clustering, and showed its usefulness on three novel tasks in concept retrieval, concept clustering, concept sentiment prediction both quantitatively and qualitatively. Our key findings are the following:

We showed that visual sentiment concepts from multiple languages can be effectively represented in a common semantic space via machine translation or multilingual semantic alignment, building on recent advances in distributional semantics. The proposed approach based on a skip-gram model trained on real-world image metadata from Flickr, with a summed combination of concept word embeddings or by direct concept learning, achieved superior performance than alternatives. This enabled multilingual clustering of visual sentiment concepts in 11 languages, allowed us to better hierarchically organize MVSO ontology [1], and provided a deep multilingual perspective into portrait imagery. On sentiment prediction, the proposed multimodal features achieved superior performance than text-only and image-only features. Moreover, the aligned concept representations provide superior performance to translation-based representations on all three examined tasks.

We also performed clustering of multilingual concepts represented by word embeddings using $k$-means. We found that when we use embeddings that are closer to the domain of the concepts (with *Flickr* corpora), closer to the original language of the concept [33], and learned using ANPs as single tokens (Section 6), we were able to achieve higher semantic consistency within clusters. However, the highest sentiment consistency for the full MVSO corpus was achieved by the largest embedding dataset on news domain [4], which included global news and possibly covered multicultural representation. This suggests that understanding sentiment requires better coverage of universal concepts. However, given that the original text document data of Google News Corpus is not publicly available [43], we could not test how it would perform by learning with ANPs as single tokens, and also analyze how much of the original documents cover culturally universal concepts. As future work, we will explore more comprehensive evalution methods for clustering multilingual visual sentiment concepts, given that our consistency metrics favor higher number of clusters.

Owing to this framework, we believe we have introduced opportunities toward addressing at least two problems: visual sentiment prediction and multilingual concept discovery and analysis, in both generic imagery and for images from the portrait vertical. Below, we highlight some interesting observations and findings:

- **The importance of multiple modalities.** For both applications, we exploited powerful tools from both visual and the textual analysis. In our visual sentiment prediction experiments, we found that, by combining visual and textual features together, we can improve the accuracy of a sentiment classifier. Moreover, we found that only when using visual features to describe sentiment we can actually expose cross-cultural differences between sentiment models. Visual analysis was useful in clustering discovery as well: only through our pure visual consistency metric we could find language-specific visual patterns related to multilingually aligned concepts, thus enabling a deep exploration of our dataset. Finally, we would not have been able to perform portrait-only analysis if we did not have a computer vision-based face detector to retain face-related concepts only.
- **The importance of cross-cultural analytics for sentiment analysis.** Similar to previous work [1], we found that language-specific sentiment detectors are crucial to build accurate sentiment models that take into account users from different cultures and languages. We also found that there exist profound differences among language communities in the perception and depection of visual sentiment concepts.
- **Cross-cultural analytics: Eastern vs. Western.** Consistently across our analysis, we found that Eastern language communities and Western language communities tend to use different ways to visualise sentiment. We found in Section 7 that sentiment models of Latin languages tend to be similar, i.e. with high cross-cultural prediction accuracy, but that their cross-cultural prediction performances would drop when testing on Chinese languages. In cluster discovery, we found that Western and Eastern languages tend to assign different sentiment values to the some visual concepts. Finally, in our portrait analysis, we clustered languages according to the concepts they tend to attach to portrait images, and found a clear separation between Eastern and Western languages.
- **Cross-cultural analytics: Universal vs. Encultured.** In Sections 8 and 9 we exposed differences and commonalities across languages. In terms of sentiment, we found

that all languages agree on giving high sentiment to images related to happy children, young children, warm hearts: these seem to be universally positive concepts. We also found that, consistently across languages, concepts attached to images with faces tend to show higher sentiment than all other concepts, and the bigger the face, the higher the sentiment. In contrast, concepts evoking different meanings in different languages, such as *abandoned house* or *historical building* tend to have higher sentiment diversity across cultures.

– **Cross-cultural analytics: Visual vs. Sentiment.** In our cluster discovery analysis, we used a visual consistency metric to discover visual, non-alignable differences in multilingual clusters. An interesting finding was that the strongly positive and negative concepts tend to have consistent visual representations across cultures than weakly positive and negative ones. However, culture-specific concepts such as beauty or religious and historic aspects show that distinct visual patterns for different languages.

There are several interesting future work directions that can be pursued based on this study. From a modeling perspective, the proposed concept embeddings in multiple languages could be further improved by directly learning them from the concept co-occurrence statistics (Section 4.3.2), for instance, by modeling the proposed visual semantic distance using both visual and textual features. The multimodal representations would potentially capture richer contextual information as compared to embeddings learned from a single modality. Another interesting direction is to learn multimodal concept embeddings that are sentiment-biased directly from the visual concept sentiment data. Such concept embeddings may provide further improvements on the visual sentiment concept retrieval, clustering and sentiment prediction. From a analysis perspective, our framework can support research related to culture, psycholinguistics, and human behavior analysis at the individual level or community levels. Lastly, there are several real-world applications which can benefit advertising and e-commerce through culture-awareness using the proposed framework such as multicultural image clustering, multicultural image sentiment analysis and multicultural image query expansion, as demonstrated in [40].

## References


1. B. Jou, T. Chen, N. Pappas, M. Redi, M. Topkara, and S.-F. Chang, "Visual affect around the world: A large-scale multilingual visual sentiment ontology," in *ACM International Conference on Multimedia*, (Brisbane, Australia), pp. 159–168, 2015.
2. J. Turian, L. Ratinov, and Y. Bengio, "Word representations: A simple and general method for semi-supervised learning," in *48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, (Uppsala, Sweden), pp. 384–394, 2010.
3. R. Collobert, J. Weston, L. Bottou, M. Karlen, K. Kavukcuoglu, and P. Kuksa, "Natural language processing (almost) from scratch," *Journal of Machine Learning Research*, vol. 12, pp. 2493–2537, 2011.
4. T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, vol. abs/1301.3781, 2013.
5. J. Pennington, R. Socher, and C. D. Manning, "GloVe: Global vectors for word representation," in *Empirical Methods in Natural Language Processing*, pp. 1532–1543, 2014.
6. R. Al-Rfou, B. Perozzi, and S. Skiena, "Polyglot: Distributed word representations for multilingual NLP," *CoRR*, vol. abs/1307.1662, 2013.
7. A. Klementiev, I. Titov, and B. Bhattarai, "Inducing crosslingual distributed representations of words," in *Proceedings of COLING 2012*, (Mumbai, India), pp. 1459–1474, 2012.
8. W. Y. Zou, R. Socher, D. Cer, and C. D. Manning, "Bilingual word embeddings for phrase-based machine translation," in *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, (Seattle, WA, USA), pp. 1393–1398, 2013.
9. K. M. Hermann and P. Blunsom, "Multilingual models for compositional distributed semantics," in *Annual Meeting of the Association for Computational Linguistics*, (Baltimore, Maryland), pp. 58–68, 2014.
10. A. P. S. Chandar, S. Lauly, H. Larochelle, M. M. Khapra, B. Ravindran, V. C. Raykar, and A. Saha, "An autoencoder approach to learning bilingual word representations," *CoRR*, vol. abs/1402.1454, 2014.
11. F. Hill, R. Reichart, and A. Korhonen, "Simlex-999: Evaluating semantic models with (genuine) similarity estimation," *CoRR*, vol. abs/1408.3456, 2014.
12. E. Bruni, N. K. Tran, and M. Baroni, "Multimodal distributional semantics," *Journal of Artificial Intelligence Research*, vol. 49, pp. 1–47, Jan. 2014.
13. C. Silberer and M. Lapata, "Learning grounded meaning representations with autoencoders," in *52nd Annual Meeting of the Association for Computational Linguistics*, (Baltimore, Maryland), pp. 721–732, June 2014.
14. A. Lazaridou, N. T. Pham, and M. Baroni, "Combining language and vision with a multimodal skip-gram model," in *Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, (Denver, Colorado), pp. 153–163, 2015.
15. A. Karpathy, A. Joulin, and F. Li, "Deep fragment embeddings for bidirectional image sentence mapping," in *Advances in Neural Information Processing Systems 27*, pp. 1889–1897, Curran Associates, Inc., 2014.
16. R. Kiros, R. Salakhutdinov, and R. S. Zemel, "Unifying visual-semantic embeddings with multimodal neural language models," *CoRR*, vol. abs/1411.2539, 2014.
17. M. Faruqui, and C. Dyer. "Improving vector space word representations using multilingual correlation." Association for Computational Linguistics, 2014.
18. R. Socher, A. Karpathy, Q. V. Le, C. D. Manning, and A. Y. Ng, "Grounded compositional semantics for finding and describing images with sentences," *TACL*, vol. 2, pp. 207–218, 2014.
19. J. Mao, W. Xu, Y. Yang, J. Wang, and A. L. Yuille, "Explain images with multimodal recurrent neural networks," *CoRR*, vol. abs/1410.1090, 2014.
20. S. Kottur, R. Vedantam, J. M. F. Moura, and D. Parikh, "Visual word2vec (vis-w2v): Learning visually grounded word embeddings using abstract scenes," *CoRR*, vol. abs/1511.07067, 2015.
21. T. Schnabel, I. Labutov, D. Mimno, and T. Joachims, "Evaluation methods for unsupervised word embeddings," in *Conference on Empirical Methods in Natural Language Processing*, (Lisbon, Portugal), pp. 298–307, 2015.

22. O. Levy, Y. Goldberg, and I. Dagan, "Improving distributional similarity with lessons learned from word embeddings," *Transactions of Association for Computational Linguistics*, vol. 3, pp. 211–225, 2015.
23. T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, pp. 3111–3119, 2013.
24. R. Lebret and R. Collobert, "Word embeddings through hellinger pca," in *Conference of the European Chapter of the Association for Computational Linguistics*, (Gothenburg, Sweden), pp. 482–490, 2014.
25. M. Baroni and R. Zamparelli, "Nouns are vectors, adjectives are matrices: Representing adjective-noun constructions in semantic space," in *Conference on Empirical Methods in Natural Language Processing*, (Cambridge, MA, USA), pp. 1183–1193, 2010.
26. R. Socher, B. Huval, C. D. Manning, and A. Y. Ng, "Semantic compositionality through recursive matrix-vector spaces," in *Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, (Jeju Island, Korea), pp. 1201–1211, 2012.
27. H. Schmid, "Probabilistic part-of-speech tagging using decision trees," in *International Conference on New Methods in Language Processing*, (Manchester, UK), 1994.
28. W. A. Freiwald and D. Y. Tsao, "Neurons that keep a straight face," *National Academy of Sciences*, vol. 111, no. 22, pp. 7894–7895, 2014.
29. M. Redi, N. Rasiwasia, G. Aggarwal, and A. Jaimes, "The beauty of capturing faces: Rating the quality of digital portraits," in *IEEE International Conference and Workshops on Automatic Face and Gesture Recognition*, (Ljubljana, Slovenia), pp. 1–8, 2015.
30. B. Jou, S. Bhattacharya, and S.-F. Chang, "Predicting viewer perceived emotions in animated GIFs," in *ACM International Conference on Multimedia*, (Orlando, Florida, USA), pp. 213–216, 2014.
31. S. Bakhshi, D. A. Shamma, and E. Gilbert, "Faces engage us: Photos with faces attract more likes and comments on instagram," in *ACM Conference on Human Factors in Computing Systems*, (Toronto, ON, Canada), pp. 965–974, 2014.
32. S. Liao, A. K. Jain, and S. Z. Li, "A fast and accurate unconstrained face detector," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, pp. 211–223, 2016.
33. Ammar, Waleed, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith."Massively multilingual word embeddings," arXiv preprint arXiv:1602.01925, 2016.
34. Uwe Quasthoff, Matthias Richter, Christian Biemann, "Corpus Portal for Search in Monolingual Corpora, Proceedings of the fifth international conference on Language Resources and Evaluation," LREC, pp. 1799-1802, Genoa, 2006.
35. Nikolaos Pappas, Miriam Redi, Mercan Topkara, Brendan Jou, Hongyi Liu, Tao Chen, and Shih-Fu Chang, "Multilingual visual sentiment concept matching," In ACM International Conference on Multimedia Retrieval, pp. 151–158, New York, USA, 2015.
36. Bo Pang and Lillian Lee, "Seeing stars: Exploiting class relationships for sentiment categorization with respect to rating scales," in 43rd Annual Meeting on Association for Computational Linguistics, pp. 115–124, Ann Arbor, Michigan, 2005.
37. Jou, Brendan, and Shih-Fu Chang. "Deep Cross Residual Learning for Multitask Visual Recognition." In Proceedings of the 2016 ACM Conference on Multimedia Conference, pp. 998–1007, Amsterdam, Netherlands, 2016.
38. Pang, Bo, and Lillian Lee, "Opinion mining and sentiment analysis," Foundations and trends in information retrieval 2, no. 1-2 (2008): 1-135.
39. Pang, Bo, Lillian Lee, and Shivakumar Vaithyanathan, "Thumbs up?: sentiment classification using machine learning techniques," in ACL-02 conference on Empirical methods in natural language processing-Volume 10, pp. 79-86, Philadelphia, PA, 2002.
40. Hongyi Liu, Brendan Jou, Tao Chen, Mercan Topkara, Nikolaos Pappas, Miriam Redi, and Shih-Fu Chang, "Complura: Exploring and leveraging a large-scale multilingual visual sentiment ontology," pp .417–420, New York, USA, 2015.
41. Turney, Peter D, "Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews," in 40th Annual Meeting on Association for Computational Linguistics, pp. 417-424, Philadelphia, PA, 2002.
42. Maas, Andrew L., Raymond E. Daly, Peter T. Pham, Dan Huang, Andrew Y. Ng, and Christopher Potts, "Learning word vectors for sentiment analysis," in 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1, pp. 142-150, 2011.
43. Mikolov, Tomas, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean, "Distributed representations of words and phrases and their compositionality," In Advances in neural information processing systems, pp. 3111-3119. 2013.
44. Tang, Duyu, Furu Wei, Nan Yang, Ming Zhou, Ting Liu, and Bing Qin, "Learning Sentiment-Specific Word Embedding for Twitter Sentiment Classification," in 52nd Annual Meeting of the Association for Computational Linguistics, pp. 1555-1565, Baltimore, MD, 2014.
45. Hu, Minqing, and Bing Liu, "Mining and summarizing customer reviews," In 10th ACM SIGKDD international conference on Knowledge discovery and data mining, pp. 168-177, Seattle, WA, 2004.
46. Zhuang, Li, Feng Jing, and Xiao-Yan Zhu, "Movie review mining and summarization," in 15th ACM international conference on Information and knowledge management, pp. 43-50, Arlington, VA, 2006.
47. Titov, Ivan, and Ryan McDonald, "Modeling online reviews with multi-grain topic models," in 17th international conference on World Wide Web, pp. 111-120, Beijing, China, 2008.
48. Sauper, Christina, Aria Haghighi, and Regina Barzilay, "Incorporating content structure into text analysis applications," in 2010 Conference on Empirical Methods in Natural Language Processing, pp. 377-387, Cambridge, MA, 2010.
49. Lu, Bin, Myle Ott, Claire Cardie, and Benjamin K. Tsou, "Multi-aspect sentiment analysis with topic models," in 2011 IEEE 11th International Conference on Data Mining Workshops, pp. 81-88, Washington, DC, 2011.
50. McAuley, Julian, Jure Leskovec, and Dan Jurafsky, "Learning attitudes and attributes from multi-aspect reviews," in 2012 IEEE 12th International Conference on Data Mining, pp. 1020-1025, Brussels, Belgium, 2012.
51. Pappas, Nikolaos, and Andrei Popescu-Belis, "Explaining the stars: Weighted multiple-instance learning for aspect-based sentiment analysis," in Conference on Empirical Methods in Natural Language Processing, pp. 455-466, Doha, Qatar, 2014.
52. Louis-Philippe Morency, Rada Mihalcea, and Payal Doshi, "Towards multimodal sentiment analysis: Harvesting opinions from the web," in 13th international Conference on Multimodal Interfaces, pp. 169-176, Tokyo, Japan, 2011.
53. Veronica Rosas, Rada Mihalcea, and Louis-Philippe Morency, "Multimodal sentiment analysis of Spanish online videos," in IEEE Intelligent Systems 28, no. 3: 38-45, 2013
54. Erik Cambria, Bjorn Schuller, Yunqing Xia, and Catherine Havasi, "New avenues in opinion mining and sentiment analysis," in IEEE Intelligent Systems 28, no. 2: 15-21, 2013.
55. Borth, Damian, Rongrong Ji, Tao Chen, Thomas Breuel, and Shih-Fu Chang, "Large-scale visual sentiment ontology and detectors using adjective noun pairs," in 21st ACM international conference on Multimedia, pp. 223-232, Barcelona, Spain, 2013.
56. Quanzeng You, Jiebo Luo, Hailin Jin, and Jianchao Yang, "Cross-modality Consistent Regression for Joint Visual-Textual Sentiment Analysis of Social Multimedia," In 9th ACM International

Conference on Web Search and Data Mining, pp. 13-22, San Fransisco, USA, 2016.
57. Soujanya Poria, Erik Cambria, and Alexander Gelbukh, “Deep convolutional neural network textual features and multiple kernel learning for utterance-level multimodal sentiment analysis,” In 2015 Conference on Empirical Methods in Natural Language Processing, pp. 2539-2544, Lisbon, Portugal, 2015.
58. Dodds, Peter Sheridan, Eric M. Clark, Suma Desu, Morgan R. Frank, Andrew J. Reagan, Jake Ryland Williams, Lewis Mitchell et al. “Human language reveals a universal positivity bias.” Proceedings of the National Academy of Sciences 112, no. 8 (2015): 2389-2394.
59. Soujanya Poria, Erik Cambria, Newton Howard, Guang-Bin Huang, and Amir Hussain.“Fusing audio, visual and textual clues for sentiment analysis from multimodal content.” Neurocomputing 174: 50-59, 2016.
60. Li, Hongzhi, Joseph G. Ellis, Heng Ji, and Shih-Fu Chang. “Event specific multimodal pattern mining for knowledge base construction.” In Proceedings of the 2016 ACM on Multimedia Conference, pp. 821-830. ACM, 2016.