

# Multilingual Text-Image Olfactory Object Matching Based on Object Detection

Yi Shao<sup>1</sup>, Yang Zhang<sup>1</sup>, Wenbo Wan<sup>1</sup>, Jing Li<sup>1,\*</sup> and Jiande Sun<sup>1,\*</sup>

<sup>1</sup>Shandong Normal University, China

## Abstract

Among the emerging multimodal tasks, the study of olfactory modality is very classic. Detecting olfactory objects in text-image data currently faces a challenge that there is no good standard way to uniformly represent text elements and image elements that elicit the same smell, especially in multilingual data. In addition, this problem also faces the problem of imbalance in the number of positive and negative samples. Therefore, this paper proposes a method based on object detection, constructs a unified text-image object representation method based on olfactory information, and alleviates the negative impact of sample imbalance to a certain extent. The overall performance of the proposed method on the four languages is close to that of the SOTA method.

## 1. Introduction

As a kind of sensory information that can directly mobilize human memory and stimulate human emotions, olfactory information has a huge potential utilization value similar to visual information [1]. In the past two decades, with a "sensory revolution" [2] [3], researchers' perspective has gradually shifted from text and image content to information of more sensory dimensions. However, due to the scarcity of related terms in the field of smell, and the lack of dedicated direct representation method for olfactory features [4] comparing with text and image data with mature feature extraction methods, extracting olfactory information from text and image data is an ongoing research problem.

In this paper, we explore the MUSTI task of MediaEval2022 [5]. In the MUSTI task, subtask 1 is required to detect whether the image and text in each sample of the development set contain objects that cause the same olfactory experience, and subtask 2 further requires to point out what these objects are. We construct a method for matching olfactory information in text-image data based on object detection. Object detection usually also faces the problem of imbalance in the number of positive and negative samples. This type of method can not only detect olfactory objects from images, but also alleviate the problem of sample imbalance to a certain extent. We also constructed a list of objects that cause similar olfactory experience, called the "approximate object list", as a way of expressing olfactory information, and established the connection between similar objects in multiple languages.

## 2. Related Work

Although there have been studies related to odor in the field of machine learning, most of them focus on the chemical molecular structure that causes odor [6] [7] [8], and there is still a lack

---


*MediaEval'22: Multimedia Evaluation Workshop, January 13–15, 2023, Bergen, Norway and Online*

\*Corresponding author.

✉ 2021020981@stu.sdnu.edu.cn (Y. Shao); 2021317099@stu.sdnu.edu.cn (Y. Zhang); wanwenbo@sdnu.edu.cn (W. Wan); lijingjdsun@hotmail.com (J. Li); jiandesun@hotmail.com (J. Sun)

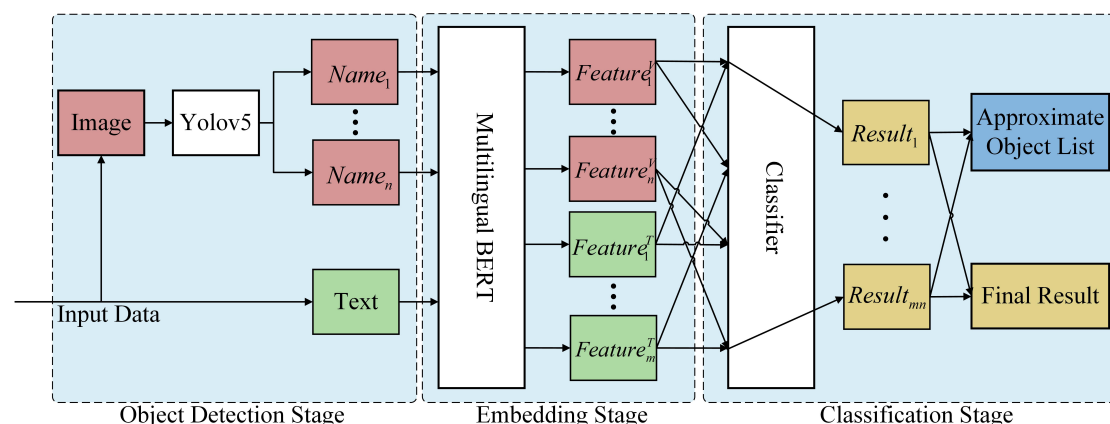


© 2022 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

of research on using natural language processing methods and computer vision methods to explore olfactory information. Menini et al. [9] propose a method that focuses on olfactory events rather than traditional single terms, and constructs a multilingual olfactory benchmark. Menini et al. [10] constructed a taxonomy of olfactory-related terms using WordNet [11] and Google n-grams, and enriched the terms with temporal information, making it possible to trace the relative usage of these odors over the past centuries. Tonelli et al. [12] capture odor events and situations in text, and manually mark the main actors in the scene, and make some modifications to the olfactory-related frame and annotation practice based on FrameNet [13].

### 3. Approach



**Figure 1:** The overview of proposed method.

#### 3.1. Object Detection Stage

An overview of the proposed method flow is shown in Figure 1. First, we performed Yolov5-adapted annotations on the images in the development set. Specifically, we counted all object names contained in the subtask 2 labels of all samples in the development set, performed cross-lingual deduplication, and finally manually annotated these objects on the image. Then we used the labeled images to fine-tune the pretrained Yolov5 model [14]. The pre-trained Yolov5 model obtained after fine-tuning can be obtained the *textual* names of the objects in the image as visual olfactory objects  $O^V$ . For each input sample, its image will be input to the pre-trained Yolov5, and all the unique object names  $O^V$  in the image will be output. The object names in the images are textual, so they can be extracted as embedding vectors by BERT along with the text.

#### 3.2. Embedding Stage

We use the multilingual BERT, a text feature extraction model that can be applied to English, French, German, and Italian. Since the visual olfactory objects are represented by discrete words, and the embedding vector of each word does not contain context information, we also input the sample text into BERT in the form of *discrete* words. The visual and textual object embedding features obtained by BERT are expressed as  $F^V$  and  $F^T$ .

### 3.3. Classification Stage

After embedding stage, we can directly get the matching result of the same objects in the image and text. But we still need to judge the olfactory similarity between different olfactory objects, because highly similar olfactory properties may appear between them. To this end, we design an olfactory similarity classifier  $C$ . Specifically, the inputs  $F_i^V$  and  $F_j^T$  will go through a fully connected layer and a cosine similarity calculation, and then  $C$  will output the matching result according to the final similarity result  $S$ .

When the  $S$  value exceeds a threshold  $s$ , the olfactory objects  $O_i^V$  and  $O_j^T$  corresponding to  $F_i^V$  and  $F_j^T$  are considered to cause the same olfactory experience, and then  $O_i^V$  and  $O_j^T$  will be added to subtask 2 result list and an approximate object list. Finally, according to whether the length of result list of subtask 2 is 0, the result of subtask 1 is obtained. By the approximate object list, not only can a connection be established between different objects that cause similar olfactory experience, but also a bridge can be established between different language expressions of the same object, thereby achieving multilingual adaptability during model inference.

## 4. Results and Analysis

### 4.1. Models Performance Comparison under Data Imbalance

There are 1789 negative samples and 601 positive samples in the development set data, and the ratio is roughly equal to 3:1. We tried to directly use ResNet101 and BERT to extract visual and text features respectively and splicing them for binary classification prediction, but the performance was low. After that, we changed the loss function to Focal Loss [15] that adapts to the unbalanced samples, and the effect improvement is not significant. Finally, we tried the method based on object detection. As shown in Figure 1, the proposed object detection-based method significantly outperforms the ResNet and Focal Loss-based methods on a small number of positive samples, which proves that it is feasible to use object detection methods to solve sample imbalance.

**Table 1**

Performance of different methods for data imbalance problem on development set

Methods	Negative Samples			Positive Samples			avg
	Precision	Recall	F1-score	Precision	Recall	F1-score	F1-score
ResNet101 + BERT + Entropy Loss	0.8625	0.7667	0.8118	0.4864	0.6271	0.5362	0.6740
ResNet101 + BERT + Focal Loss	0.8356	0.8472	0.8414	0.5133	0.4915	0.5022	0.6718
<b>ours</b>	<b>0.8743</b>	<b>0.8694</b>	<b>0.8719</b>	<b>0.6083</b>	<b>0.6186</b>	<b>0.6134</b>	<b>0.7427</b>

### 4.2. Baseline comparison experiment

The performance comparison among the proposed method and baseline methods proposed in [16] on the final official test set is shown in Table 2. It can be seen that the proposed model is close to the best baseline in overall performance.

As shown in Table 2, the subtask 1 performance of the proposed method performs well on English and Italian data, but poorly on French and German data. We speculate that this is because the number of samples in different languages varies greatly (en:800, de:482, fr:304, it:804), resulting in different distributions of visual objects corresponding to different languages.

**Table 2**

Performance Comparison with Baseline Models on test set

Subtask	Models	F1-score				
		en	de	fr	it	avg
Subtask 1	dummy baseline	0.4285	0.4289	0.3333	0.4273	0.4075
	mUniter finetuned	0.4473	0.4644	0.3605	0.5020	0.4473
	mUniter-MUSTI	0.6965	0.4579	0.5022	0.6535	0.6011
	mUniter-SNLI-MUSTI	0.7482	<b>0.5014</b>	<b>0.5053</b>	0.6850	<b>0.6176</b>
	<b>ours</b>	<b>0.7867</b>	0.4568	0.3743	<b>0.7501</b>	0.6033
Subtask 2	<b>ours</b>	0.7427	0.7276	0.4599	0.7487	0.6708

That is, many visual-olfactory objects in the German and French example images were not fully learned by Yolov5.

In order to confirm this conclusion, we input the images corresponding to each language sample into Yolov5 again, and made a macro average of the output visual and olfactory objects. The results are shown in Table 3, which verified our conjecture.

**Table 3**

F1-score macro avg of multilingual visual olfactory objects

Language	en	de	fr	it
F1-score avg of visual objects	0.5495	0.4050	0.2041	0.5439

According to Table 2, the subtask 2 performance of the proposed method is equally good on German data besides English and Italian. This is because the German dataset has more images containing only a few olfactory objects. The fewer olfactory objects in the image, the smaller the probability that the image will be correctly detected by Yolov5, and the F1 score of the corresponding sample in subtask 1 is often lower. In other words, we not only need to perform cross-lingual data enhancement, but we also need to perform data enhancement for different visual objects to obtain a better Yolov5.

## 5. Discussion and Outlook

We build a method for detecting whether images and text elicit similar olfactory experiences, with overall performance approaching that of SOTA methods. But it still has some disadvantages.

For the relationship between approximate olfactory objects, the method proposed in this paper only considers the most basic synonym relationship, but in fact we can use them to build complex graph structures to represent the associations between different olfactory objects. In addition, we still need to perform data enhancement for olfactory objects and different languages to improve the performance of Yolov5, which is the core of the model.

## Acknowledgement

Thanks to the organizers of the MediaEval2022, especially to those organizers for MUSTI. This work was supported in part by the Scientific Research Leader Studio of Jinan (Grant No. 2021GXRC081), and in part by the Joint Project for Smart Computing of Shandong Natural Science Foundation (Grant No. ZR2021LZH010, ZR2020LZH015, and ZR2022LZH012).

## References

- [1] P. Lisena, D. Schwabe, M. van Erp, R. Troncy, W. Tullett, I. Leemans, L. Marx, S. C. Ehrich, Capturing the semantics of smell: The odeuropa data model for olfactory heritage information, in: *European Semantic Web Conference*, Springer, 2022, pp. 387–405.
- [2] D. Howes, *Charting the sensorial revolution*, 2006.
- [3] C. Classen, Other ways to wisdom: Learning through the senses across cultures, *International Review of Education* 45 (1999) 269–280.
- [4] B. Winter, Synaesthetic metaphors are neither synaesthetic nor metaphorical, *Perception metaphors* 19 (2019) 105–126.
- [5] A. Hürriyetoğlu, T. Paccosi, S. Menini, M. Zinnen, P. Lisena, K. Akdemir, R. Troncy, M. van Erp, MUSTI - Multimodal Understanding of Smells in Texts and Images at MediaEval 2022, in: *Proceedings of MediaEval 2022 CEUR Workshop*, 2022.
- [6] C. C. Licon, G. Bosc, M. Sabri, M. Mantel, A. Fournel, C. Bushdid, J. Golebiowski, C. Robardet, M. Plantevit, M. Kaytoue, et al., Chemical features mining provides new descriptive structure-odor relationships, *PLoS computational biology* 15 (2019) e1006945.
- [7] B. Sanchez-Lengeling, J. N. Wei, B. K. Lee, R. C. Gerkin, A. Aspuru-Guzik, A. B. Wiltschko, Machine learning for scent: Learning generalizable perceptual representations of small molecules, *arXiv preprint arXiv:1910.10685* (2019).
- [8] D. Wu, D. Luo, K.-Y. Wong, K. Hung, Pop-cnn: Predicting odor pleasantness with convolutional neural network, *IEEE Sensors Journal* 19 (2019) 11337–11345.
- [9] S. Menini, T. Paccosi, S. Tonelli, M. Van Erp, I. Leemans, P. Lisena, R. Troncy, W. Tullett, A. Hürriyetoğlu, G. Dijkstra, et al., A multilingual benchmark to capture olfactory situations over time, in: *Proceedings of the 3rd Workshop on Computational Approaches to Historical Language Change*, 2022, pp. 1–10.
- [10] S. Menini, T. Paccosi, S. S. Tekiroğlu, S. Tonelli, Building a multilingual taxonomy of olfactory terms with timestamps, in: *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, 2022, pp. 4030–4039.
- [11] G. A. Miller, Wordnet: a lexical database for english, *Communications of the ACM* 38 (1995) 39–41.
- [12] S. Tonelli, S. Menini, Framenet-like annotation of olfactory information in texts, in: *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, 2021, pp. 11–20.
- [13] J. Ruppenhofer, M. Ellsworth, M. Schwarzer-Petruck, C. R. Johnson, J. Scheffczyk, *FrameNet II: Extended theory and practice*, Technical Report, International Computer Science Institute, 2016.
- [14] ultralytics, yolov5, <https://github.com/ultralytics/yolov5>, 2020.
- [15] T.-Y. Lin, P. Goyal, R. Girshick, K. He, P. Dollár, Focal loss for dense object detection, in: *Proceedings of the IEEE international conference on computer vision*, 2017, pp. 2980–2988.
- [16] K. Akdemir, A. Hürriyetoğlu, R. Troncy, T. Paccosi, S. Menini, M. Zinnen, V. Christlein, Multimodal and Multilingual Understanding of Smells using ViLBERT and mUNITER, in: *Proceedings of MediaEval 2022 CEUR Workshop*, 2022.