Research Report - Summary of the paper

Recognizing Food Places in Egocentric Photo-Streams Using MACNet & Self-Attention Mechanism

Artificial Vision & Pattern Recognition

Màster en Enginyeria de la Seguretat Informàtica i Intel·ligència Artificial (URV)
Màster interuniversitari en Intel·ligència Artificial (UPC, UB & URV)

Student: Óscar Cubeles Ollé

Abstract

This report provides a summary of the paper *"Recognizing Food Places in Egocentric Photo-Streams Using MACNet & Self-Attention Mechanism"*. The paper presents a novel deep learning model, **MACNet+SA**, designed to classify food-related environments in egocentric photo-streams captured throughout daily activities. By combining the **MACNet architecture** with a **self-attention mechanism**, the proposed model improves the classification of sequential images, extracting both spatial and temporal information. This report outlines the key methodologies, results, and future work discussed in the paper, emphasizing the potential applications of the model in dietary tracking and health monitoring.

28th January 2026

1 Introduction

Obesity and overweight are major risk factors for chronic diseases including cancer, diabetes, and cardiovascular diseases. Developing automated systems to analyze daily lifestyle patterns particularly nutrition-related behaviors represents a promising healthcare approach. Understanding metrics such as duration, location, and social context of food-related activities is crucial for monitoring and improving people's eating habits. In this context, wearable cameras offer a unique opportunity to capture these patterns through egocentric photo streams, automatically documenting users' environments, object interactions, and visited places. By analyzing food place classifications and time spent in these locations, insights to help individuals improve their dietary behaviors can be provided.

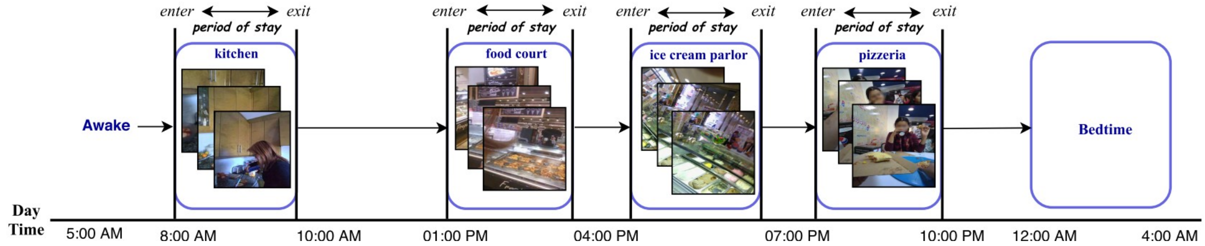


Figure 1: Example of Photostream sequence [1]

The main contributions of this paper are: (1) Design and development of MACNet+SA, a novel attention-based deep network combining multi-scale Atrous convolutional networks with self-attention mechanisms to improve food place classification accuracy, and (2) Application of the MACNet+SA model for analyzing sequences of images to detect food-related events and behaviors. This approach leverages the EgoFoodPlaces dataset, which contains images from multiple individuals using life-logging cameras.

2 Relative Work

Traditional scene classification approaches include discriminative methods (logistic regression, SVMs, boosting) and generative hierarchical Bayesian models that capture complex scene relationships. The initial CNN breakthroughs using AlexNet, ResNet with skip connections, and VGG, combined with datasets like Places2 and SUN397, have significantly outperformed traditional methods. However, scene recognition faces unique challenges compared to object recognition, particularly due to the diverse environments surrounding people. Egocentric images present additional challenges due to the vast variety of real-world food place environments and the wide range of capture perspectives from a person's viewpoint. The authors previously introduced MACNet, based on atrous CNNs, for food place classification using pre-trained ResNet on individual images without temporal dependencies. To the best of our knowledge, this is the first work on food place pattern classification based on analyzing events from streams of egocentric images.

3 Proposed Approach

RNNs and attention-based models are commonly used to process sequential data, especially in tasks like image or video captioning and sentiment analysis. They are useful when the input has a temporal structure, such as egocentric photo-streams. Attention mechanisms can be hard or soft. Hard attention focuses only on part of the input, while soft attention assigns weights to all inputs using a softmax function. A popular type of soft attention is self-attention, where the model learns which parts of the representation are more important by comparing elements within the same input.

In this paper, self-attention is implemented with LSTM cells to capture temporal dependencies and compute attention scores across the photo-stream. The proposed network has three main parts: feature extraction with MACNet, an LSTM-based attention module, and a final prediction module for food place recognition.

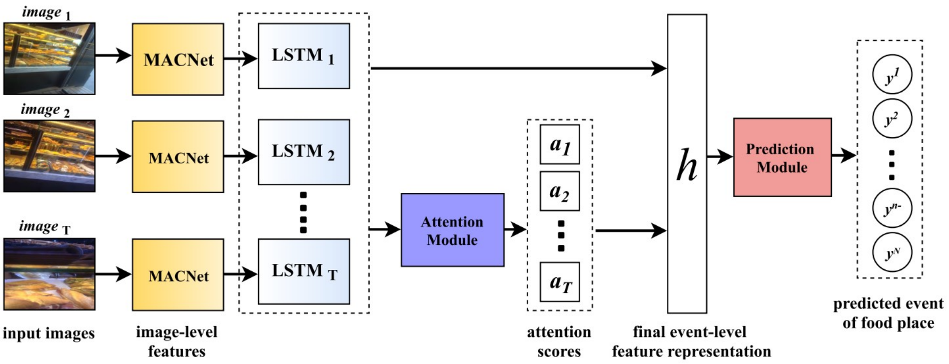


Figure 2: Proposed architecture [1]

3.1 MACNet: Multi-Scale Atrous Convolutional Network

In MACNet, the input image is resized to five different resolutions, where each scale is half the size of the previous one. The original image has a resolution of 224×224, which is the standard input size for ResNet. Each of these scaled images is then processed by a block of an atrous (dilated) CNN using three different dilation rates (1, 2, and 3). This allows the network to capture visual information at multiple spatial scales. Then, four pretrained ResNet-101 layers are used sequentially to extract feature maps at different levels of abstraction.

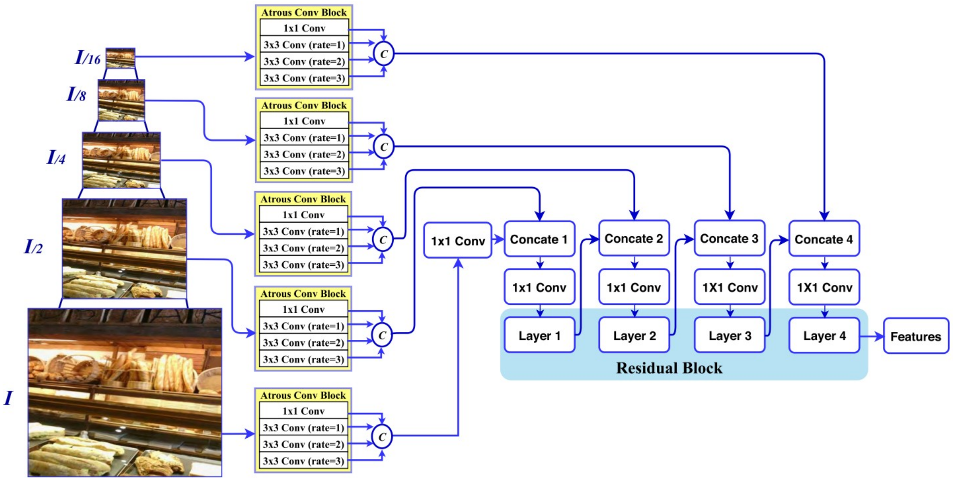


Figure 3: Proposed MACNet architecture [1]

At each stage, the feature map produced by the atrous CNN is concatenated with the output of the corresponding ResNet block, and passed to the next stage. The sequence starts with the original image features from the Atrous block and concatenated with the output of the half size image features from the Atrous block and so on. Finally, the features from the fourth (last) layer are used as the final representation of the input image.

3.2 LSTM Cells

An LSTM, or *Long Short-Term Memory* cell, is a fundamental building block of **MACNet-SA**, which is designed to handle sequential data while avoiding the vanishing gradient problem that standard RNNs often face.

The core idea of an LSTM is its gated structure, which carefully controls the flow of information:

- The **forget gate** decides which past information to discard.

- The **input gate** decides what new information to add to the memory.
- The **cell state** acts as the long-term memory, carrying relevant information across time steps.
- Finally, the **output gate** determines what information to send to the next layer.

These gates allow the network to remember important information over long sequences while ignoring irrelevant details.

In the context of this paper, LSTM cells are used to generate attention-based features, meaning the network can focus on the most relevant parts of the input sequence. By combining memory from LSTMs with attention mechanisms, the model captures both long-term dependencies and the dynamic importance of different inputs, improving the feature representation for the task at hand.

3.3 Self Attention Mechanism

After extracting features from each image in an event using MACNet, these image-level features $(x_0, x_1, \ldots, x_T)$ are fed sequentially into a set of LSTM cells $(\text{LSTM}_1, \ldots, \text{LSTM}_T)$. Each LSTM processes the sequence of images, capturing contextual dependencies. This means that the output of each LSTM cell $(\text{LSTM}_t \in R^H)$ not only represents the current image but also encodes information from the previous images, allowing the network to understand temporal or sequential patterns across the event.

However, not all images in an event contribute equally to understanding the overall event. To address this, they introduce an attention module. Each LSTM output (LSTM_t) is compared with a global attention vector $(V \in R^H)$, which is learned during training. The dot product between the LSTM output and this attention vector gives a score for each image:

$$S_t = \langle V, \text{LSTM}_t \rangle$$

reflecting its importance. These scores are then normalized using a softmax function to produce attention weights:

$$\alpha_t = \frac{\exp(S_t)}{\sum_{t=1}^{T} \exp(S_t)}$$

ensuring that the weights sum to one.

Using these attention weights (α_t), to compute a weighted average of the LSTM outputs:

$$h = \sum_{t=1}^{T} \alpha_t \, \text{LSTM}_t$$

Images with higher attention weights contribute more to the final vector, while less relevant images are down-weighted. The result is a single event-level feature vector (h) that summarizes all the image-level features, capturing both the sequential context provided by the LSTM and the relative importance of each image determined by attention.

This event-level representation (h) is then used to train the prediction module, enabling the network to accurately identify the event associated with the sequence of images. In essence, the combination of LSTM memory and attention weighting allows the model to focus on what matters most, while still considering the full context of the image sequence.

3.4 Prediction Module

There are various types of prediction modules available in the literature. In this work, a fully connected neural network is used as a multi-label event prediction module. Given the event-level feature vector h, obtained from the LSTM and attention mechanisms, the probability of the n-th event is computed as:

$$\hat{y}_n = p(y_n \mid h) = \frac{1}{1 + e^{-(w_n h + b_n)}} \in [0, 1], \tag{1}$$

where $\hat{y}_n$ denotes the predicted label of the n-th event, y_n is the corresponding ground-truth label, $n = 1, \ldots, N$, and N is the total number of event classes. The parameters w_n and b_n represent the weight vector and bias term of the classifier associated with the n-th event, respectively.

Since multiple events can occur simultaneously, the task is formulated as a multi-label classification problem. The entire model is trained end-to-end by minimizing the multi-label classification loss defined as:

$$\ell = -\frac{1}{N} \sum_{n=1}^{N} E(y_n, \hat{y}_n), \tag{2}$$

where $E(\cdot)$ denotes the cross-entropy loss function.

4 Experimental Results

4.1 Dataset Used

The paper uses the EgoFoodPlaces dataset from previous work, extending it with additional images. The dataset consists of 16 different locations captured using a lifelogging camera worn on the user's chest, covering the entire day from morning to night. Instead of still images, each class contains a set of events, with one event every 10 seconds and each event consisting of 10 frames (e.g., 10 minutes of recording yields 60 events or 600 seconds). The dataset presents several challenges, including motion blur, occlusions from body parts or objects, and black images. To address these, the applied preprocessing was: Blurriness detection: The variance of the Laplacian was computed for each image. Images with a variance below 500 were considered blurry, while higher variance indicated a sharp image. Removal of low-information images: K-Means clustering with K=3was applied to the pixels of each image. If a single cluster contained more than 90

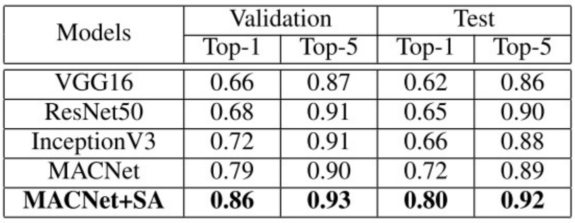
The dataset is unbalanced, reflecting real-life frequency of visits to different locations—for example, the supermarket class has more instances than the butcher's shop. The class labels match those in the Places2 dataset, totaling 22 classes. Finally, the dataset was split into 70

4.2 Experimental Setup

The model was implemented in PyTorch and trained using the Adam optimizer with a learning rate of 0.001 and a step value of 20. The LSTM consists of 6 layers, each with a hidden size of 2048 and a dropout rate of 0.3. The self-attention mechanism has 22 layers, corresponding to the 22 classes in the dataset. Data augmentation was applied to improve generalization, including random cropping, adjustments to brightness and contrast, as well as rotation and translation. The model was trained with a batch size of 64 over 100 epochs. Training was performed on an NVIDIA GTX1080 GPU with 11GB of memory, taking approximately one day to complete.

4.3 Evaluation

The performance of MACNet+SA was evaluated against several CNN architectures, including VGG16, ResNet50, InceptionV3, and MACNet without self-attention (SA), using F1 score, Top-1, and Top-5 accuracy as metrics. MACNet+SA achieved the highest average F1 scores, reaching 0.86 on the validation set and 0.80 on the test set, outperforming all other networks in most classes. MACNet without SA showed a behavior similar to InceptionV3 in terms of average F1. Overall, the ranking in terms of F1 score is MACNet+SA first, followed by MACNet and then InceptionV3. Adding self-attention improved the performance of MACNet by approximately 7–8%. However, for certain classes such as butcher's shop, dining room, and both indoor and outdoor markets, MACNet without SA performed slightly better, likely because these scenes can be effectively recognized from single images without the need for sequential modeling.



Models	Validation		Test	
	Top-1	Top-5	Top-1	Top-5
VGG16	0.66	0.87	0.62	0.86
ResNet50	0.68	0.91	0.65	0.90
InceptionV3	0.72	0.91	0.66	0.88
MACNet	0.79	0.90	0.72	0.89
MACNet+SA	**0.86**	**0.93**	**0.80**	**0.92**

Figure 4: Average Top-1 and Top-5 classification accuracy of the different networks for the task [1]

In terms of Top-1 and Top-5 accuracy, MACNet+SA again achieved the best results on both validation and test sets, followed by MACNet. Despite its overall strong performance, the network still exhibits misclassifications, particularly for classes with visually or contextually similar scenes. For example, on the validation set, 17.78% of fast-food restaurant events were misclassified as restaurant, 22.80% of picnic area events were misclassified as outdoor market, and ice cream parlour events were misclassified 22% as supermarket and 14% as outdoor market. Some delicatessen and candy store events were also misclassified as supermarket. On the test set, a substantial portion of events were misclassified as restaurant, including 36.74% of fast-food restaurant events, 33.01% of banquet hall events, 33.01% of picnic area events, 34.06% of beer hall events, and 18.49% of bar events. These errors typically occur in scenes that share similar contextual or visual features, making them challenging even for human observers. In summary, the addition of self-attention improves MACNet's ability to recognize sequential patterns, leading to better overall performance, especially for classes that require sequential image understanding. Nevertheless, there is still room for improvement in differentiating classes with visually or contextually similar scenes.

5 Conclusions

This work presents MACNet+SA, a deep learning system for food place classification from egocentric photo-streams captured throughout daily activities. The primary motivation behind this approach is to enable the automatic generation of dietary reports, which can be used to analyze eating habits and support healthier lifestyle choices. By recognizing food-related environments over time, the system provides contextual information that complements traditional food intake analysis, offering a more comprehensive understanding of users' dietary behavior. The proposed model extends the original MACNet architecture by incorporating temporal modeling and self-attention, allowing it to process sequences of images rather than isolated frames. Image-level features are extracted using atrous convolutions, while LSTM cells with a self-attention mechanism effectively capture temporal dependencies across events. Experimental results on the EgoFoodPlaces dataset demonstrate that MACNet+SA consistently outperforms state-of-the-art methods, including VGG16, ResNet50, InceptionV3, and MACNet without self-attention. The model achieves strong performance in terms of F1 score, Top-1, and Top-5 accuracy on both validation and test sets, confirming the benefit of incorporating sequential information and attention mechanisms for egocentric scene understanding. Looking ahead, their future work will focus on translating this system into a real-world mobile application that integrates an egocentric camera with a personal mobile device. Such an application would enable on-device or online processing to generate personalized dietary reports while minimizing privacy risks by avoiding the storage of raw images. This direction opens the door to practical, privacy-aware tools for monitoring eating behavior and supporting long-term healthy dietary routines.

References

[1] Sarker, M. M. K., Rashwan, H. A., Akram, F., Talavera, E., Banu, S. F., Radeva, P., and Puig, D. (2019). *Recognizing Food Places in Egocentric Photo-Streams Using Multi-Scale Atrous Convolutional Networks and Self-Attention Mechanism.* IEEE Access.