Research Project Proposal: A Mobile Visual Language Model for Real-Time Personalized Nutrition Assistance

This project proposes a mobile application that integrates Visual Language Models (VLMs), computer vision, and human-computer interaction (HCI) to provide real-time, personalized nutrition assistance. The app will calculate a user's Body Mass Index (BMI) from visual inputs, estimate nutritional needs, and guide users in food selection and portion control using multimodal feedback (voice, text, visual). By focusing on edge computing efficiency, explainability, and dataset creation, this project introduces novel contributions to dietary assessment tools, targeting human-level nutritional intuition with practical deployment on mobile devices.

Accurate dietary assessment is critical for health management, yet existing methods often rely on manual logging or memory, lacking real-time feedback. Mobile computer vision-based applications for food recognition exist but suffer from limitations such as poor differentiation between food and non-food items and minimal explainability. This project aims to address these gaps by developing a mobile app that:

- Uses VLMs and computer vision to analyze food and body metrics.
- Provides intuitive, real-time guidance on food choices and quantities.
- Operates efficiently on resource-constrained mobile devices.

Problem Statement: There is a need for an innovative, user-friendly, and efficient nutrition assistance tool that leverages modern AI to provide personalized, real-time dietary recommendations on mobile platforms.

Literature Review

Current research on mobile food recognition includes:

- Food Recognition: Tools like Food-101 dataset-based models use convolutional neural networks (e.g., YOLO) for identifying food items but struggle with non-food differentiation.
- Volume and Calorie Estimation: Some apps estimate calories from images, but accuracy is limited, and explanations for decisions are rare.
- Datasets: Available datasets (e.g., Food-101, UNIMIB2016) focus on raw or specific foods, lacking diversity for cooked dishes or contextual settings like canteens.
- Edge Computing: Lightweight models (e.g., MobileNet) enable on-device inference, but large VLMs remain challenging to deploy efficiently.

Gaps

- Lack of real-time, multimodal feedback.
- Limited personalization and explainability.
- Insufficient datasets for specific contexts (e.g., canteen dishes).
- Real time voice/ text based assistant system

Relevant Technologies:

- VLMs (e.g., CLIP) for combined image-text understanding.
- YOLO for real-time object segmentation.
- OpenFoodFacts API for nutritional data.

Proposed Solution

The proposed mobile app will:

1. Accept user height and weight inputs.
2. Use visual analysis to calculate BMI and nutritional needs.
3. Provide real-time food analysis (shape, size, quantity) via camera input.
4. Offer multimodal guidance (voice, text, visual) during food selection.
5. Assess portion sizes post-selection and advise on adequacy.
6. Suggest food choices from lunch boxes or based on eating patterns.
7. Recalculate BMI weekly, adapting recommendations over time.

Technical Approach

- VLM: A lightweight VLM (e.g., a distilled version of CLIP) will analyze food images and contextualize them with nutritional data.
- Computer Vision: YOLOv5 (nano version) for real-time food segmentation, combined with depth estimation for volume analysis.
- HCI: Multimodal interface with voice (e.g., Google Speech API), text, and visual overlays for user interaction.
- Edge Computing: Model compression (e.g., quantization, pruning) to ensure efficient on-device performance.
- Data: Leverage Food-101 and OpenFoodFacts API, supplemented by a custom canteen dish dataset. If that's not enough, collect our own data or generate synthetic data for diversification.

Key Features

- Real-Time Assistance: Immediate feedback during food selection without need to take pictures.
- Personalization: Adapts to user eating patterns using simple machine learning updates.
- Intuition Goal: Estimates object weight (e.g., an apple at 200 grams) with 80-90% accuracy using size and shape cues and based on nutritional data of foods or fruits it compares with the weight and predicts the amount of food quantifiers such as vitamin, minerals, carbs, fats, and so on.

Novelty and Innovations

This project introduces several innovative elements:

1. Edge Computing Efficiency: Optimize a VLM for mobile devices using lightweight frameworks (e.g., TensorFlow Lite), ensuring low latency and battery usage.
2. Explainable AI: Provide reasons for recommendations (e.g., “This portion exceeds your caloric goal by 200 kcal”), enhancing user trust.
3. Multimodal Interaction: Combine voice, text, and visual feedback for a seamless experience, rare in existing tools.
4. Dataset Creation: Develop a canteen-specific dataset, addressing a gap in contextual food recognition.
5. Synthetic Data: Use AI-generated content (AIGC) for edge cases (e.g., rotten apples), improving robustness.

6. Volume-to-Weight Estimation: Innovate weight prediction (e.g., apples) using 2D image cues and statistical priors, avoiding complex 3D hardware.

Proposed Methodology

Phase 1: Data Collection

- Task 0: Evaluate datasets (Food-101, UNIMIB2016). If insufficient, proceed to dataset creation.
- Task 1: Create a canteen dish dataset:
  - Capture images of common canteen foods (e.g., rice, curry).
  - Annotate with nutritional info from restaurant menus or OpenFoodFacts.
  - Generate synthetic data for edge cases using AIGC tools (e.g., DALL-E).

Phase 2: Model Development

- VLM: Fine-tune a lightweight CLIP variant on food images and text descriptions.
- Computer Vision: Implement YOLOv5-nano for segmentation and a regression model for volume/weight estimation.
- Edge Optimization: Use quantization and pruning to reduce model size to <50 MB, targeting <1-second inference on mid-range phones.

Phase 3: HCI Design

- Design an intuitive interface with:
  - Voice commands (e.g., “Analyze my plate”).
  - Visual overlays highlighting food items.
  - Text feedback on portion adequacy.

Phase 4: Evaluation

- Metrics:
  - Accuracy: Food recognition (>90%), weight estimation (>80%).
  - Efficiency: Inference time (<1s), battery impact (<5% per use).
  - User Satisfaction: Surveys on usability and trust.
- Testing: Conduct trials in a university canteen setting.

Challenges and Solutions

- Dataset Availability: If no suitable dataset exists, creating one is feasible but time-intensive. Start small with canteen dishes.
- Scaling: Begin with raw and simple cooked foods, expanding later with more data.
- Cooked Food Complexity: Use menu-based nutritional priors and focus on visual cues (e.g., portion size) rather than exact ingredients.

Expected Contributions

- A practical mobile app for real-time nutrition assistance.
- A new canteen dish dataset, valuable for research.
- Advances in lightweight VLM deployment on edge devices.
- Improved explainability and user interaction in dietary tools.

References

- Food-101 Dataset: https://data.vision.ee.ethz.ch/cvl/datasets_extra/food-101/
- UNIMIB2016: https://mldta.com/dataset/unimib2016-food-database/
- OpenFoodFacts API: https://world.openfoodfacts.org/