Segmenting Egocentric Videos to Highlight Personal Locations of Interest

Antonino Furnari, Giovanni Maria Farinella, Sebastiano Battiato
University of Catania - Department of Mathematics and Computer Science
{*furnari,gfarinella,battiato*}*@dmi.unict.it*

Introduction and Motivations

With the increasing availability of wearable cameras, the acquisition of egocentric videos is becoming common in many scenarios including law enforcement, assistive technologies, and life-logging. However, the absence of explicit structure in such videos (e.g., video chapters), makes their exploitation difficult. Depending on the considered goal, long egocentric videos tend to contain much uninformative content like for instance transiting through a corridor, walking, or driving to the office. Therefore, automated tools are needed to enable faster access to the information stored in such videos and index their visual content. Towards this direction, researches have investigated methods to produce short informative video summaries, recognize the actions performed by the wearer, and segment the videos according to detected ego-motion patterns.

While current literature focuses on providing general-purpose methods which are usually optimized on data acquired by many users, we argue that, given the subjective nature of egocentric videos, more attention should be devoted to user-specific methods. More specifically, we propose to segment unstructured egocentric videos into coherent shots related to user-specified personal locations of interest. We consider a personal location as: *a fixed, distinguishable spatial environment in which the user can perform one or more activities which may or may not be specific to the considered location*. According to this notion, a personal location is specified at the instance level (e.g., my kitchen, my office, my car), rather than at the category level (e.g., a kitchen, an office, a car). Given a set of personal locations of interest to be considered for the segmentation of egocentric sequences, the task is to understand (for every frame in the video) if it is related to either one of the considered personal locations of interest or none of them (in which case it will be referred to as a negative sample to be rejected).

Figure 1 shows a schema of the investigated problem: given an input video and minimal user-specified training data (i.e., short video-clips of the personal locations of interest), the system should be able to segment the video highlighting the presence of the considered locations of interest as well as rejecting the negative frames. In a real-world scenario in which the system is set up by the end user himself, training must be simple and achievable with few training data. Moreover, given the large variability exhibited by egocentric videos, it is unfeasible to ask the user to acquire a significant quantity of negative samples. Therefore, we assume that only positive samples are provided by the user while no negative sample is required at training time. In particular, for each location of interest, the user is only asked to acquire a short video ($\approx$ 10 seconds) looking around in order to capture the most likely views. To support experiments, we collected a dataset of egocentric videos related to 10 different personal locations, plus various negative ones. The considered locations arise from a possible daily routine: Car, Coffee Vending Machine (C.V.M.), Office, Lab Office, TV, Piano, Kitchen, Sink, Studio, Garage. The training set comprises a short 10-second video-clip for each location of interest. A validation set containing one longer video for each location is provided (validation videos are about 5 minutes long). We also provide 10 sequences comprising different changes of locations to test segmentation accuracy. Given only positive training data, the proposed method is able to discriminate among user-specified locations of interest, reject negative samples and segment the input video.

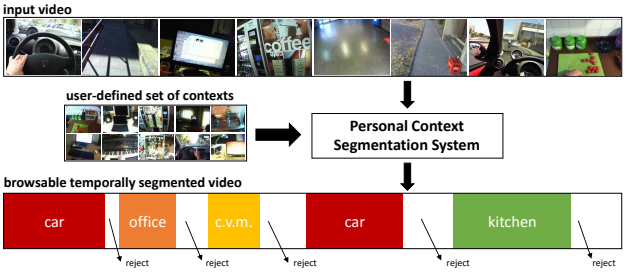

Figure 1. Overall scheme of the proposed temporal segmentation of an egocentric video.

Method

Let $\mathcal{V} = \{I_1, \ldots, I_n\}$ be an egocentric video composed by frames I_i. Our system must be able to 1) correctly classify each frame I_i as one of the user-specified locations, 2) reject

		Accuracy			Comp. Perf.	
Id	Settings	Discrim.	+Rej.	+HMM	Dim.	Time
$[a]$		76.90	69.60	73.83	378 MB	13.23 ms
$[b]$	L	83.30	76.06	83.22	378 MB	13.13 ms
$[c]$	L ND	**94.53**	**85.00**	**88.63**	378 MB	13.10 ms
$[d]$	L 128	83.07	77.49	82.84	34 MB	10.32 ms
$[e]$	L ND 128	77.09	71.99	73.59	34 MB	10.28 ms
$[f]$	L LR	**92.31**	**81.00**	**85.37**	26 MB	10.23 ms
$[g]$	L ND NE	73.84	76.42	79.69	378 MB	12.82 ms
$[h]$	SVM [1]	87.76	74.14	79.64	423 MB	97.83 ms

Table 1. Methods $[a]$ - $[g]$ are obtained fine-tuning VGG on the training set using the architectural settings specified in the second column: L the convolutional layers are locked, ND dropout is disabled, 128 fully connected layers are reduced to 128 units and reinitialized, LR fully connected layers are replaced by a single logistic regression layer, NE the model is trained on both positive and negative samples. Method $[h]$ consists of a cascade of SVM classifiers trained on the features extracted using the VGG-S network as detailed in [1]. Reported times are average per-image processing times. They include rejection but do not include HMM-related computations. Maxima per column are reported in **underlined bold digits**, while second maxima are reported in **bold digits**.

negative frames (i.e., images not related to any of the considered locations) and 3) segment temporally coherent subsequences related to the locations of interest. The system eventually returns the segmentation $\mathcal{S} = \{C_1, \ldots, C_n\}$, where $C_i \in \{0, \ldots, M-1\}$ is the class label associated to frame I_i ($C_i = 0$ representing the negative class label) and $M = 11$ in our case (10 locations, plus the negative class).

Discrimination among positive frames is obtained fine-tuning the VGG Convolutional Neural Network (CNN) pre-trained on ImageNet, on the training set comprising only positive samples. Since the discriminative model ignores the presence of negative frames, it only allows to estimate the posterior probability:

$$p(C_i | I_i, C_i \neq 0). \tag{1}$$

We propose to quantify the probability of a given frame I_i to be a negative sample ($p(C_i = 0|I_i)$), as the uncertainty of the discriminative model in predicting class labels of last k observed frames (we use $k = 30$). As proposed in [2], model uncertainty is estimated computing the variation ratio of the distribution of last k labels $\mathcal{Y}_i^k = \{y_i, \ldots, y_k\}$ predicted by maximizing the posterior probability reported in Eq. (1). We finally assign the probability of I_i being a negative as follows:

$$p(C_i = 0|I_i) = 1 - \frac{\sum \mathbb{1}(y_i = \tilde{\mathcal{Y}}_i^k)}{|\mathcal{Y}_i^k|} \tag{2}$$

Considering that $C_i = 0$ and $C_i \neq 0$ are disjoint events (and hence $p(C_i \neq 0|I_i) = 1 - p(C_i = 0|I_i)$), the probabilities reported in Eq. (1) and (2) are combined as:

$$p(C_i|I_i) = \begin{cases} p(C_i = 0|I_i) & \text{if } C_i = 0 \\ p(C_i \neq 0|I_i) \cdot p(C_i|I_i, C_i \neq 0) & \text{otherwise} \end{cases}. \tag{3}$$

To enforce temporal coherence between subsequent predictions, we consider a Hidden Markov Model (HMM) with 11 hidden states. We define the state transition probability as in [3] in order to discourage sudden label changes:

$$p(C_i|C_{i-1}) = \begin{cases} \varepsilon, & \text{if } C_i \neq C_{i-1} \\ 1 - (M-1)\varepsilon, & \text{otherwise} \end{cases} \tag{4}$$

where ε is a small constant. The final segmentation of the input egocentric video is obtained using the Viterbi algorithm:

$$\mathcal{S}^* = \arg\max_{\mathcal{S}} p(\mathcal{S}|\mathcal{V}). \tag{5}$$

Results

Table 1 summarizes the experimental results. In columns 3 to 5, we assess the impact of the proposed negative rejection method and Hidden Markov Model component on the accuracy of the overall system (measured as the portion of correctly labeled frames). When rejection is not considered (i.e., results reported in the Discrim. - Discrimination - column), negative samples are removed from the test set.

We show that fine-tuning a CNN with a small training set (≈ 200 samples per class in our experiments) is not trivial and many architectural details can be tuned. In particular we investigate the influence of: locking convolutional layers, disabling dropout, reducing the dimensionality of fully connected layers, and removing fully connected layers. It should be noted that some of these architectural changes positively impact computational performances both in terms of required memory and time.

We compare our method with respect to the method proposed in [1], which uses a one-class SVM classifier to reject negative samples, and the baseline $[g]$ which is not provided with a rejection mechanism and explicitly trains a negative class using negative samples (test and training negatives are separated). While all methods takes advantage of the proposed negative rejection + HMM paradigm, methods $[c]$ and $[f]$ show the highest accuracies. In particular, method $[f]$ reaches high accuracy with reduced computational requirements. It should be noted that, while discriminating among different locations is an easier task on which high accuracies are reached by the considered methods, the need of a rejection mechanism makes the problem harder. HMM usually helps reducing the gap between discrimination and discrimination+rejection performances. Proposed dataset, a video demo of the proposed system and supplementary material are available at our webpage *http://iplab.dmi.unict.it/PersonalLocations/segmentation/*

References

[1] A. Furnari, G. M. Farinella, and S. Battiato. Recognizing personal contexts from egocentric images. In *Workshop on Assistive Computer Vision and Robotics (ACVR) in conjunction with ICCV*, 2015.

[2] Y. Gal and Z. Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. *arXiv preprint arXiv:1506.02142*, 2015.

[3] R. Templeman, M. Korayem, D. Crandall, and K. Apu. PlaceAvoider: Steering First-Person Cameras away from Sensitive Spaces. In *Annual Network and Distributed System Security Symposium*, pages 23–26, 2014.