Pattern Recognition 72 (2017) 207–218

Contents lists available at ScienceDirect

Pattern Recognition

journal homepage: www.elsevier.com/locate/patcog

Organizing egocentric videos of daily living activities

Alessandro Ortis[a], Giovanni M. Farinella[a,*], Valeria D'Amico[b], Luca Addesso[b], Giovanni Torrisi[b], Sebastiano Battiato[a]

[a] *Image Processing Laboratory, Dipartimento di Matematica e Informatica, Universitã degli studi di Catania, Viale A. Doria 6, Catania - 95125 Italy*
[b] *TIM - Telecom, JOL WAVE, Viale A. Doria 6, Catania - 95125 Italy*

ARTICLE INFO

Article history:
Received 4 October 2016
Revised 3 May 2017
Accepted 7 July 2017
Available online 14 July 2017

Keywords:
First person vision
Video summarization
Video indexing

ABSTRACT

Egocentric videos are becoming popular since the possibility to observe the scene flow from the user's point of view (First Person Vision). Among the different applications of egocentric vision is the daily living monitoring of a user wearing the camera. We propose a system able to automatically organize egocentric videos acquired by the user over different days. Through an unsupervised temporal segmentation, each egocentric video is divided in chapters by considering the visual content. The obtained video segments related to the different days are hence connected according to the scene context in which the user acts. Experiments on a challenging egocentric video dataset demonstrate the effectiveness of the proposed approach that outperforms with a good margin the state of the art in accuracy and computational time.

1. Introduction and motivations

In the last years there has been a rapid emerging of wearable devices, including body sensors, smart clothing and wearable cameras. These technologies can have a significant impact on our lives if the acquired data are considered to assist the users in tasks related to the monitoring of the quality of life [1–4]. In particular, egocentric cameras enabled the design and development of useful systems that can be organized into three general categories with respect to the assistive tasks:

(i) *User Aware* - systems able to understand what the user is doing and what/how he interact with, by recognizing actions and behaviours from first person perspective.
(ii) *(ii) Environment/Objects Aware* - systems able to understand what objects surround the user, where they are with respect the user's perspective and what the environment looks like.
(iii) *Target Aware* - systems able to understand what others are doing, and how they interact with the user that is wearing the device.

The egocentric monitoring of a person's daily activities can help to stimulate the memory of users that suffer from memory disorders [5]. Several works on recognition and indexing of daily living activities of patients with dementia have been recently proposed [6–9]. The exploitation of aids for people with memory problems is proved to be one of the most effective ways to aid rehabilitation [10]. Furthermore, the recording and organization of daily habits performed by a patient can help a doctor to have a better opinion with respect to the specific patient's behaviour and hence his health needs. To this aim, a set of egocentric videos recorded among different days with a camera hold by a patient can be analysed by experts to monitor the user's daily living activities for assistive purposes. The live recording for life logging applications poses challenges on how to perform automatic index and summarization of big personal multimedia data [11].

Beside assistive technologies, the segmentation and semantic organization of egocentric videos is useful in many application scenarios where wearable cameras have recently become popular, including lifelogging [11], law enforcement [12] and social cameras [13]. Other applications of egocentric video analysis are related to action and activity recognition from egocentric videos [14–16], and recognition of interaction-level activities from videos in first-person view [17] (e.g., recognition of human-robot interactions). For all these applications, the segmentation of daily egocentric videos into meaningful chapters and the semantic organization of such video segments related to different days is an important first step which adds structure to egocentric videos, allowing tools for the indexing, browsing and summarization of egocentric video sets.

In the last years, several papers have addressed different problems related to vision tasks from first person perspective. The work in [18] proposes a temporal segmentation method of ego-

* Corresponding author.
E-mail addresses: ortis@dmi.unict.it (A. Ortis), gfarinella@dmi.unict.it (G.M. Farinella), valeria1.damico@telecomitalia.it (V. D'Amico), luca.addesso@telecomitalia.it (L. Addesso), giovanni.torrisi@telecomitalia.it (G. Torrisi), battiato@dmi.unict.it (S. Battiato).

http://dx.doi.org/10.1016/j.patcog.2017.07.010
0031-3203/

centric videos with respect to 12 different activities organized hierarchically upon cues based on wearer's motion (e.g., static, sitting, standing, walking, etc.). A benchmark study considering different wearable devices for context recognition with a rejection mechanism is presented. The system discussed in [19] aims at segmenting unstructured egocentric videos to highlight the presence of given personal contexts of interest. In [4] the authors perform a benchmark on the main representations and wearable devices used for the task of context recognition from egocentric videos. A method that takes a long input video and returns a set of video subshots depicting the essential moments is detailed in [20]. The method proposed in [21] learns the sequences of actions involved in a set of habitual daily activities to predict the next actions and generate notifications if there are missing actions in the sequence. The framework presented in [13] (RECfusion), is able to automatically process multiple video flows from different mobile devices to understand the most popular scenes for a group of end-users. The output is a video which represents the most popular scenes organized over time.

In this paper, we build on the RECfusion method [13] improving it for the context of daily living monitoring from egocentric videos. In RECfusion multiple videos are analysed by using two algorithms: the former is used to segment the different scenes over time (intraflow analysis). The latter is employed to perform the grouping of the videos related to the involved devices over time. As reported in the experimental results of [13], the intraflow analysis of RECfusion suffers when applied to egocentric videos because they are highly unstable due to the user's movements.

The framework proposed in this paper allows to have better performances for egocentric videos organization, on both segmentation accuracy and computational costs. The proposed method takes a set of egocentric videos regarding the daily living of a user among different days, and performs an unsupervised segmentation of them. The obtained video segments among different days are then organized by contents. The video segments of the different days sharing the same contents are then visualized by exploiting an interactive web-based user interface. In our framework we use a unique representation for the frames which is based on CNN features [22] (for both intraflow and between flows analysis) instead of the two different representations based on SIFT and color histograms as proposed in RECfusion. Experiments show that the proposed framework outperforms RECfusion for daily living egocentric video organization. Moreover, the approach obtains better accuracy than RECfusion for the popularity estimation task for which RECfusion has been designed.

The rest of the paper is organized as follows. Section 2 presents the designed framework. Section 3 describes the considered wearable dataset. The discussion on the obtained results is reported in Section 4. In Section 5 the developed system is compared with respect to RECfusion [13] on mobile videos. Finally Section 6 concludes the paper and gives insights for further works.

2. Proposed framework

The proposed framework performs two main steps on the videos acquired by a wearable camera: temporal segmentation and segment organization. Fig. 1 shows the scheme of the overall pipeline related to the our system.

Starting from a set of egocentric videos recorded among multiple days (Fig. 1(a)), the first step performs an intraflow analysis of each video to segment it with respect to the different scenes observed by the user (Fig. 1(b)). Each video is then segmented by employing temporal and visual correlations between frames of the same video (Fig. 1(b): the colour of each block identifies a scene observed within the same video). Then, the segments obtained over different days referred to the same contents are grouped by means of a between flows analysis which implements an unsupervised clustering procedure aimed to semantically associate video segments obtained from different videos (Fig. 1(c): the colour of each block now identifies video segments associated to the same visual content, observed in different days). The system hence produces sets of video clips related to each location where the user performs daily activities (e.g., the set of the clips over days where the user washes dishes in the kitchen, the set related to the activity of the user of playing piano, and so on). The clips are organized taking into account both, visual and temporal correlations. Finally, the framework provides a web based interface to enable a browsing of the organized videos (Fig. 1(d)). In the following subsections the details on the different steps involved into the pipeline are given.

2.1. Intraflow analysis

The intraflow analysis performs the unsupervised temporal segmentation of each input video, as well as associates a scene ID to each video segment. Segments with the same content have to share the same scene ID. This problem has been addressed in [13] for videos acquired with mobile devices. To better explain the problem, tn the following we focus our analysis on the issues related to the intraflow algorithm detailed in [13] when applied on first-person videos. This is useful to introduce the main problems of a classic feature based matching approach for temporal segmentation in wearable domain. Then we present our solution for the intraflow analysis. Furthermore, in Appendix A the pseudocode describing the proposed intraflow analysis approach is reported.

2.1.1. Issues related to SIFT based templates

The intraflow analysis used in [13] compares two scenes considering the number of matchings between a reference template and the frame under consideration (current frame). The scene template is a set of SIFT descriptors that must accomplish specific properties of "reliability". When the algorithm detects a sudden decrease in the number of matchings, it refreshes the reference template extracting a new set of SIFTs and splits the video. In order to detect such changes, the system computes the value of the slope in the matching function (i.e., the variation of the number of matchings in a range interval). When the slope is positive and over a threshold (which correspond to a sudden decrease of the number of matchings between the SIFT descriptors) the algorithm finds a new template (i.e., a new block). When a new template is defined, it is compared with the past templates in order to understand if it regards a new scene or it is related to a known one (backward search phase). Although this method works very well with videos acquired with mobile cameras, it presents some issues when applied on videos acquired with wearable devices. In such egocentric videos, the camera is constantly moving due to the shake induced by the natural head motion of the wearer. This causes a continuous refresh of the reference template that is not always matched with similar scenes during the backward search. Hence, the method produces an oversegmentation of the egocentric videos. Furthermore, it requires to perform several SIFT descriptor extraction and matching operations (including geometric verifications) to exclude false positive matchings. This have a negative impact on the realtime performances. The first row of Fig. 2 shows the Ground Truth segmentation of the video acquired with a wearable camera in an home environment.[1] An example of a temporal segmentation obtained with the SIFT based interflow analysis proposed in [13] on a egocentric video is reported in the second row of Fig. 2. The

[1] The video related to the example in Fig. 2 is available for visual inspection at the URL http://iplab.dmi.unict.it/dailylivingactivities/homeday.html.

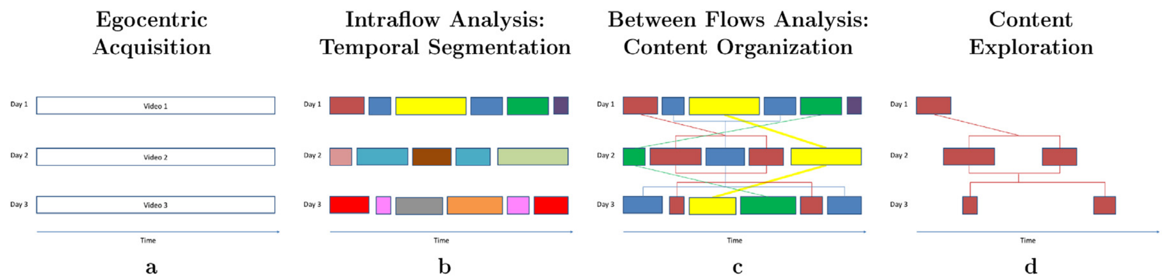

Fig. 1. Overall scheme of the proposed framework.

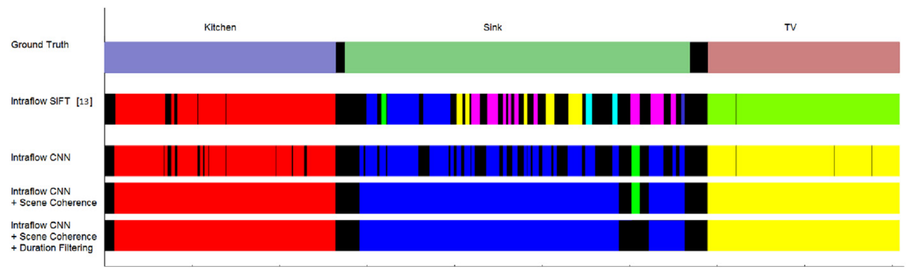

Fig. 2. Output of the intraflow analysis using SIFT and CNN features when applied to the video *Home Day 1*. The first row is the Ground Truth of the video. The second row shows the result of the intraflow analysis with the method discussed in [13]. The third row shows the result of the intraflow analysis of our method, whereas the last two rows show the results of our method with the application of the proposed scene coherence and duration filtering criteria.

algorithm works well when the scene is quite stable (e.g., when the user is watching a TV program), but it performs several errors when the scene is highly unstable due head movements. In fact, in the middle of the video related to Fig. 2, the user is washing dishes at the sink, and he is continuously moving his head. In this example the intraflow approach based on SIFT features detects a total of 8 different scenes instead of 3. The algorithm cannot find the matchings between the current frame and the reference template due to two main reasons:

1. when the video is unstable, even though the scene content doesn't change, the matchings between local features are not reliable and stable along time;
2. In a closed space such as an indoor environment, the different objects of the scene can be very close to the viewer. Hence a small movement of the user's head is enough to cause an high number of mismatches between local features.

2.1.2. CNN based image representation

To deal with the issues described in the previous section, we exploit an holistic feature to describe the whole image rather than an approach based on local features. In particular, in the intraflow analysis we represent frames by using features extracted with a Convolutional Neural Network (CNN) [23]. Specifically, we consider the CNN proposed in [22] (*AlexNet*). In our experiments, we exploit the representation obtained considering the output of the last hidden layer of *AlexNet*, which consists of a 4096 dimensional feature vector (*fc7* features). We decided to use *AlexNet* representation since it has been successfully used as a general image representation for classification purpose in the last few years [24,25]. Moreover, the features extracted by *AlexNet* have been successfully used for transfer learning [26–28]. Finally, *AlexNet* architecture is a small network compared to others (e.g., VGG [29]). Thus, it allows to perform the feature extraction very quickly. The proposed solution computes the similarity between scenes by comparing a pair of *fc7* features with the *cosine similarity* measure. The cosine similarity of two vectors measures the cosine of the angle between them. This measure is independent of the magnitude of the vectors, and is well suited to compare high dimensional sparse vectors, such as the *fc7* features. The cosine similarity of the two *fc7* feature vectors v_1 and v_2 is computed as following:

$$CosSimilarity(v_1, v_2) = \frac{v_1 \cdot v_2}{\|v_1\| \|v_2\|} \quad (1)$$

2.1.3. Proposed intraflow analysis

During the intraflow analysis, the proposed algorithm computes the cosine similarity between the current reference template and the *fc7* features extracted from the frames following the reference. When the algorithm detects a sudden decrease in the cosine similarity, it refreshes the reference template selecting a new *fc7* feature corresponding to a stable frame. As in [13], to detect such changes the system computes the value of the slope (i.e., the variation of the cosine similarity in a range interval). When the slope has a positive peak (which correspond to a sudden decrease of the cosine similarity) the algorithm finds a new template and produces a new video segment. There are two cases in which the intraflow analysis compares two templates:

1. a template is compared with the features extracted from the forward frames, when the algorithm have to check its eligibility to be a reference template for the involved scene;
2. A template is compared with the past templates during the backward checking phase to establish the scene ID.

In the first case, the elapsed time between the two compared frames depends on the sampling rate of the frames (in our experiments, we sampled at every 20 frames for videos recorded at 30 fps). Differently, the frames compared during the backward checking could be rather different due to the elapsed time between them. For this reason, when we compare a new template with a past template, we assign the templates to the same scene ID by using a weaker condition with respect to the one used in the forward verification. In the forward process, the algorithm assigns the same scene ID to the corresponding frames if the cosine similarity between their descriptors is higher than a threshold T_f (equal to 0.60 in our experiments). When the algorithm compares two templates in the backward process, it assign the same scene to the corresponding frames if the similarity is higher than a threshold T_b (equal to 0.53 in our experiments).

Besides the image representation, our intraflow algorithm introduces two additional rules:

1. In [13] each video frame is assigned to a scene ID or it is classified as Noise and hence rejected. Our approach is able to distinguish the rejected frame among the ones caused by the movement of the head of the user (Head Noise) and the frames related to the transition between two scenes in the video (Transition). When a new template is defined after a group of consecutive rejected frames, the frames belonging to the rejected group are considered as "Transition" if the duration of the block is longer than 3 s (i.e., head movements are faster than 3 s). Otherwise they are classified as "Head Noise". In case of noise, the algorithm doesn't assign a new scene ID to the frame that follows the "Head Noise" video segment because the noise is related to head movements, but the user is in the same context. When the user changes its location he changes his position in the environment. Thus, the transition between different scenes involves a longer time interval.
2. The second rule is related to the backward verification. In [13] it is performed starting from the last found template and proceeding backward. The search stops when the process finds the first past template that have a good matching score with the new template. Such approach is quite appropriate for the method in [13] because it compares sets of local features and relies on the number of achieved matchings. The approach proposed in this paper compares pairs of feature vectors instead of sets of local descriptors, and selects the past template that yields the best similarity to the new one. In particular, the method compares the new template with all the previous ones, considering all the past templates that yields a cosine similarity greater than T_b. From this set of positive cases, the algorithm selects the one that achieves the maximum similarity, even if it is not the most recent in the time.

Considering the example in Fig. 2, the segmentation results achieved with the proposed intraflow approach (third row) are much better than the ones obtained using SIFT features [13] (second row).

After the discussed intraflow analysis a segmentation refinement is performed as detailed in the following subsection.

2.1.4. Intraflow segmentation refinement

Starting from the result of the proposed intraflow analysis (see the third row of Fig. 2), we can easily distinguish between "Transition" and "Noise" blocks among all the rejected frames. A block of rejected frames is a "Noise" block if both the previous and the next detected scenes are related to the same visual content, otherwise it is marked as a "Transition block". We refer to this criteria as *Scene Coherence*. The result of this step on the example considered in Fig. 2 is shown in the fourth row. When comparing the segmentation with the Ground Truth (first row), the improvement with respect to [13] (second row) is evident. Moreover, many errors of the proposed intraflow analysis (third row) are removed. The second step of the segmentation refinement consists in considering the blocks related to user activity in a location with a duration longer than 30 s (*Duration Filtering*), unless they are related to the same scene of the previous block (i.e., after a period of noise, the scene is the same as before but it has a limited duration). We applied this criteria because, in the context of Activities of Daily Living (ADL), we are interested to detect the activities of a person in a location that have a significant duration in order to be able to observe the behavior. This refinement step follows the *Scene Coherence* one. The final result on the considered example is shown in the last row of Fig. 2. Despite some frames are incorrectly rejected (during scene changes) the proposed pipeline is much more robust than [13] (compare first, second and fifth rows of Fig. 2). This outcome is quantitatively demonstrated in the experimental section of this paper on a dataset composed by 13 egocentric videos.

2.2. Between video flows analysis

When each egocentric video is segmented, the successive challenge is to determine which video segments among the different days are related to the same content. Given a segment block b_{v_A} extracted from the egocentric video v_A, we compare it with respect to all the segment blocks extracted from the other egocentric videos v_{B_j}. To represent each segment, we consider again the CNN *fc7* features extracted from one of the frames of the video segment. This frame is selected considering the template which achieved the longer stability time during the intraflow analysis. For each block b_{v_A}, our approach assigns the ID of b_{v_A} (obtained during intraflow analysis) to all the blocks $b_{v_{B_j}}$ extracted from the other egocentric videos v_{B_j} such that

$$b_{v_{B_j}} = \underset{\bar{b}_{v_{B_j}} \in v_{B_j}}{\arg\max}\{CosSimilarity(b_{v_A}, \bar{b}_{v_{B_j}}) \mid \sigma_{(b_{v_A}, v_{B_j})} \geq T_\sigma\} \ \forall v_{B_j} \quad (2)$$

where $\sigma_{(b_{v_A}, v_{B_j})}$ is the standard deviation of the cosine similarity values obtained by considering the segment block b_{v_A} and all the blocks of v_{B_j}, and T_σ is the decision threshold. This procedure is performed for all the segment blocks b_{v_A} of the video v_A. When all the blocks of v_A have been considered, the algorithm takes into account a video of another day and the matching process between video segments of the different days is repeated until all the video segments in the pool are processed. In this way a scene ID is assigned to all the blocks of all the considered videos. The pairs of blocks with the same ID are associated to the same scene, even if they belong to different videos, and all the segments are connected in a graph with multiple connected components (as in Fig. 1(c)). When there is a high variability in the cosine similarity values (i.e., the value of $\sigma_{(b_{v_A}, v_{B_j})}$ is high), the system assigns the scene ID to the segment block that achieved the maximum similarity. When a block matches with two blocks related to two different scene IDs, the system assigns the scene ID related to the block which achieved the highest similarity value. When a block isn't matched, it means that all the similarity values of the miss-matched blocks are similar. This causes low values of σ and helps the system to understand that the searched scene ID is not present (i.e., scenes with only one instance among all the considered videos). To better explain the proposed approach, the pseudocode related to the above described between video flows analysis is reported in Appendix A.

3. Dataset

To demonstrate the effectiveness of the proposed approach, we have considered a set of representative egocentric videos to perform the experiments. The egocentric videos are related to different days and have been acquired using a Looxcie LX2 wearable camera with a resolution of 640 × 480 pixels. The duration of each video is about 10 min. The videos are related to the following scenarios:

- **Home Day:** a set of four egocentric videos taken in a home environment. In this scenario, the user performs typical home activities such as cooking, washing dishes, watching TV, and playing piano. This set of videos has been borrowed from the *10contexts* dataset proposed in [19] which is available at the following URL: http://iplab.dmi.unict.it/PersonalLocations/segmentation/.
- **Office Day:** a set of six egocentric videos taken in a home and in different office environments. This set of videos concerns several activities performed during a typical day. Also this set of videos belongs to the *10contexts* dataset [19].

Table 1
Intraflow performances on the considered egocentric video dataset obtained using [13] and the proposed approach. Each test is evaluated considering the accuracy of the temporal segmentation (Q), the computation time (Time) and the number of the scenes detected by the algorithm (Scenes). The accuracy is measured as the percentage of correctly classified frames with respect to the Ground Truth. The measured time includes the feature extraction process.

			Intraflow proposed in [13] Approach			Proposed Intraflow with Segmentation Refinement		Proposed Interflow Approach		
Video	Scenario	Scenes	Q	Scenes	Time	Q	Scenes	Q	Scenes	Time
1	HomeDay1	3	62.5%	8	20’45”	77.5%	4	**92.5%**	**3**	**1’23”**
2	HomeDay2	3	71.6%	3	20’18”	80.3%	4	**94.5%**	**3**	**1’46”**
3	HomeDay3	3	64.3%	5	19’03”	79.7%	5	**94.3%**	**3**	**1’21”**
4	HomeDay4	3	84.4%	3	8’36”	**91.8%**	**3**	85.4%	2	**36”**
5	WorkingDay1	4	95.7%	5	16’16”	98.4%	5	**99.5%**	**4**	**1’22”**
6	WorkingDay2	4	82.5%	5	15’15”	98.9%	5	**100%**	**4**	**1’08”**
7	WorkingDay3	5	98.7%	6	19’02”	99.2%	6	**99.4%**	**5**	**1’29”**
8	OfficeDay1	3	23.0%	5	24’8”	55.3%	19	**66.9%**	**2**	**2’39”**
9	OfficeDay2	2	59.7%	2	10’25”	90.0%	3	**98.7%**	**2**	**1’26”**
10	OfficeDay3	3	57.2%	4	13’28”	83.6%	10	**96.3%**	**3**	**1’49”**
11	OfficeDay4	3	52.0%	4	11’37’	79.5%	5	**84.1%**	**4**	**1’41”**
12	OfficeDay5	3	70.7%	3	8’35”	86.7%	4	**95.9%**	**3**	**1’21”**
13	OfficeDay6	3	78.8%	**3**	9’33”	61.5%	5	**94.5%**	4	**1’34”**
	Average		69.4%		15’9”	83.3%		**91.9%**		**1’31”**

- **Working Day:** a set of three videos taken in a laboratory environment. The activities performed by the user in this scenario regards reading a book, working in a laboratory, sitting in front of a computer, etc.

Each video has been manually segmented to define the blocks of frames to be detected in the intraflow analysis. Moreover, the segments have been labeled with the scene ID to build the Ground Truth for the between video analysis. The Ground Truth is used to evaluate the performances of the proposed framework. The used egocentric videos, as well as the Ground Truth, are available at the following URL: http://iplab.dmi.unict.it/dailylivingactivities/.

4. Experimental results

In this section we report the temporal segmentation and the between flows video analysis results obtained on the considered dataset.

4.1. Temporal segmentation results

Table 1 shows the performances of the proposed temporal segmentation method (see Section 2.1). We compared our solution with respect to the one adopted by RECfusion [13]. For each method we computed the quality of the segmentation as the percentage of the correctly classified frames (Q), the number of detected scenes and the computational time.[2] The proposed approach obtains strong improvements (up to over 30%) in segmentation quality with respect to RECfusion (e.g., results at eight and nine rows in Table 1). Furthermore, the application of the segmentation refinements provides improvements up to 43% in segmentation quality (results at row eight in Table 1). In the fourth row of Table 1 (related to the analysis of the video *Home Day 4*), we can observe that the application of the segmentation refinements causes a decrease in performances. This video is very short compared to the other videos of the dataset. It has a duration of just4’28” and consists of a sequence of three different scenes (piano, sink and TV). The scene blocks are correctly detected by the proposed intraflow approach, which finds exactly 3 different scenes and achieves a 91,8% of accuracy without refinement. However, the middle scene (sink) has a duration of just 22 s according to the Ground Truth used in [19], thus the refinement process rejects this block due to the application of the *Duration Filtering* criteria. Considering the mean performances (last row in Table 1) our system achieves an improvement of over 14% without segmentation refinements, with over 22% of margin after the segmentation refinement. The proposed method also reduces the computational time of more than 21 min in some cases (eighth row in Table 1). It has an average computational time saving of about 13 min with respect to the compared approach [13]. The results of Table 1 show that the application of the *Scene Coherence* and the *Duration Filtering* criteria used in the segmentation refinement step (Section 2.1.4) allows to detect the correct number of scenes.

In sum, considering the qualitative and quantitative results reported respectively in Fig. 2 and Table 1, the proposed system is demonstrated to be robust for the temporal segmentation of egocentric videos, and it provides high performances with a lower computational time with respect to [13].

4.2. Between flows video analysis results

In our experiments, all the segments related to the *Home Day* and *Working Day* scenarios have been correctly connected among the different days without errors. Fig. 3 shows two timelines related to the videos of the scenario *Working Day*, whereas Fig. 4 shows the timelines related to the scenario *Home Day*. The first timeline in Figs. 3 and 4 shows the Ground Truth labeling. In the timeline, the black blocks indicate the transition intervals (to be rejected). The second timeline shows the result obtained by our framework. In this case, the black blocks indicate the frames automatically rejected by the algorithm. In order to better asses the results obtained by the proposed system, the reader can perform a visual inspection of all the results produced by our approach at the following URL: http://iplab.dmi.unict.it/dailylivingactivities/. Through the web interface the different segments can be explored.

Differently than the *Home Day* and *Working Day* scenarios, some matching error occurred in the between flow analysis of the *Office Day* scenario (see Fig. 5). Since we are grouping video blocks by contents, to better evaluate the performance of the proposed between flow analysis we considered three quality scores usually used in clustering theory. The simplest clustering evaluation measure is the *Purity* of the clustering: each cluster of video segments obtained after the between flow analysis is assigned to the most frequent class in the cluster. The *Purity* measure is hence the mean of the number of correctly assigned video blocks within the clus-

[2] The experiments have been performed with Matlab 2015a, running on a 64-bit Windows 10 OS, on a machine equipped with an Intel i7-3537U CPU and 8GB RAM.

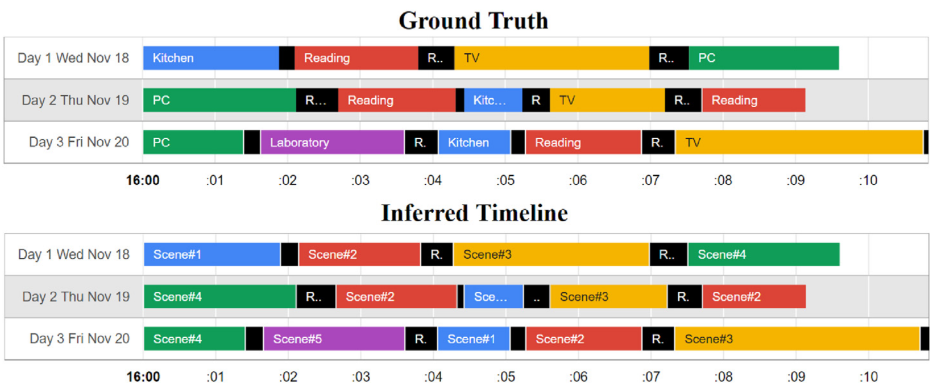

Fig. 3. The two timelines show the Ground Truth segmentation of egocentric videos related to the *Working Day* scenario and their organization (top). The inferred segmentation and organization obtained by the proposed method is reported in the bottom.

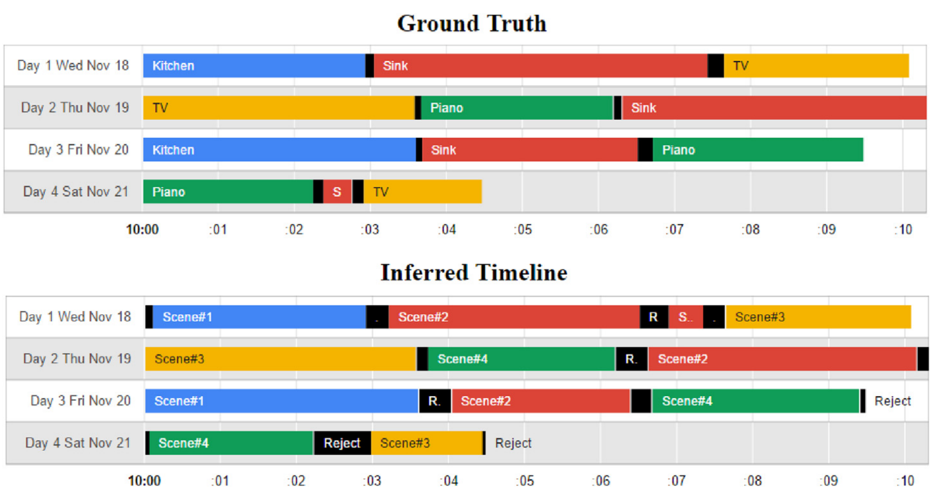

Fig. 4. The two timelines show the Ground Truth segmentation of egocentric videos related to the *Home Day* scenario and their organization (top). The inferred segmentation and organization obtained by the proposed method is reported in the bottom.

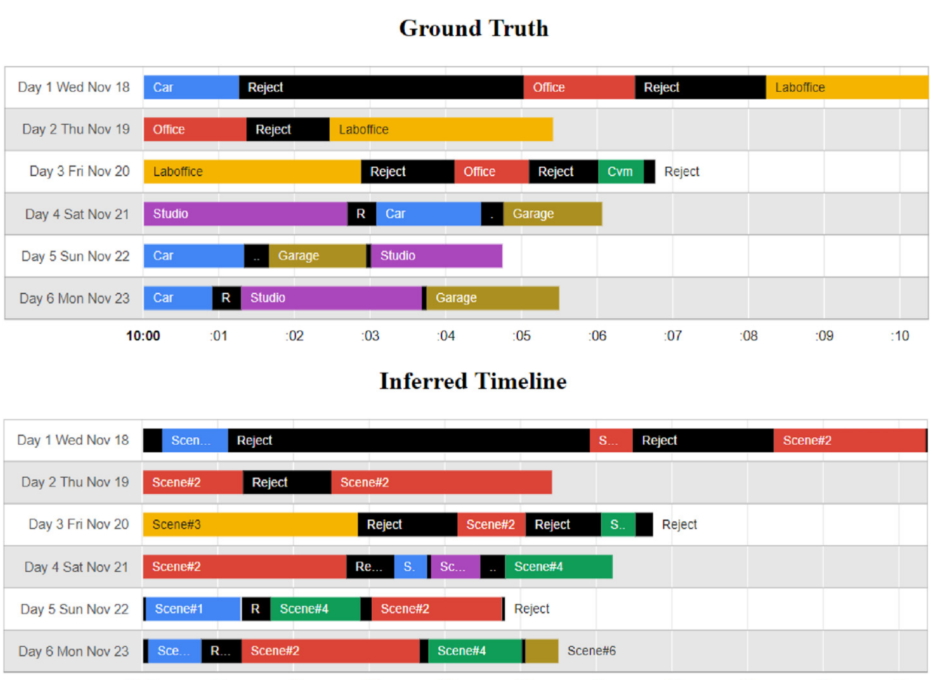

Fig. 5. The two timelines show the Ground Truth segmentation of egocentric videos related to the *Office Day* scenario and their organization (top). The inferred segmentation and organization obtained by the proposed method is reported in the bottom. In this case, the blocks related to the scenes 'office', 'studio' and 'laboffice' are clustered in the same cluster (identified by the color red), with the exception of only one 'laboffice' block.

Table 2
Between video clustering results.

Dataset	# of Videos	Original GT				New GT			
		Purity	RI	F_1	F_2	Purity	RI	F_1	F_2
WorkingDay	3	1	1	1	1	–	–	–	–
HomeDay	4	1	1	1	1	–	–	–	–
OfficeDay	6	0.68	0.81	0.49	0.57	0.95	0.89	0.80	0.75
Office+Home	10	0.58	0.83	0.41	0.54	0.74	0.87	0.61	0.67

ters.

$$Purity = \frac{1}{N}\sum_{i=1}^{k} max_j |\boldsymbol{C_i} \cap \boldsymbol{T_j}| \tag{3}$$

where N is the number of video blocks in the cluster, k is the number of clusters, $\boldsymbol{C_i}$ is the ith cluster and $\boldsymbol{T_j}$ is the set of elements of the class j that are present in the cluster $\boldsymbol{C_i}$. The clustering task (i.e., the grouping performed by the between flow analysis in our case) can be viewed as a series of decisions, one for each of the $N(N-1)/2$ pairs of the elements to be clustered [30].

The algorithm obtains a true positive decision (TP) if it assigns two videos of the same class to the same cluster, whereas a true negative decision (TN) if it assigns two videos of different class to different clusters. Similarly, a false positive decision (FP) assigns two different videos to the same cluster and a false negative decision (FN) assigns two similar videos to different clusters. With the above formalization we can compute the confusion matrix associated to the pairing task.

From the confusion matrix we compute the *Rand Index* (*RI*), which measures the percentage of the correct decisions (i.e., the pairing accuracy) of the between flow analysis:

$$RI = \frac{TP + TN}{TP + FP + FN + TN} \tag{4}$$

In order to take into account both precision and recall, we also considered the F_β measure defined as following:

$$F_\beta = (1+\beta^2)\cdot\frac{Precision \cdot Recall}{\beta^2 Precision + Recall} \tag{5}$$

Specifically, we considered the F_1 measure that weights equally precision and recall, and the F_2 measure, which weights recall higher than precision and, therefore, penalizes false negatives more than false positives. Table 2 shows the results of the proposed between flow analysis approach considering the aforementioned measures. For the *Home Day* and *Working Day* scenarios our approach achieves the maximum scores for all the considered evaluation measures (column "Original GT" of Table 2). Regarding the *Office Day* scenario, we obtain lower scores of Purity and Rand Index. The obtained values of F_1 and F_2 measures indicate that the proposed between flow approach achieves higher recall values than precision. This can be further verified by observing the Fig. 6. This figure shows the co-occurrence matrix obtained considering the *Office Day* scenario. The columns of the matrix represent the computed graph components (clusters) obtained with the between flow analysis, whereas each row represents a scene class according to the Ground Truth. The values of this matrix express the percentage of video blocks belonging to a specific class that are assigned to each graph component (according to the Ground Truth used in [19]). This figure shows that even if different blocks are included in the same graph component (FP), the majority of the blocks belonging to the same class are assigned to the same graph component (TP). The second column of the co-occurrence matrix in Fig. 6 shows that the "laboffice", "office" and "studio" blocks have been assigned to the graph component C2. This error is due to strong ambiguity in the visual content of these three classes. Fig. 8 shows some example of the frames belonging to these classes. We

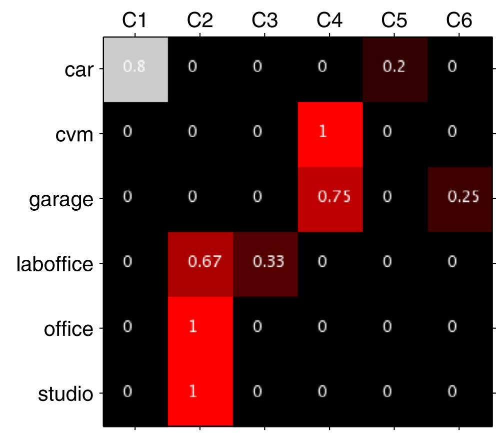

Fig. 6. Co-occurrence matrix obtained considering the *Office Day* scenario.

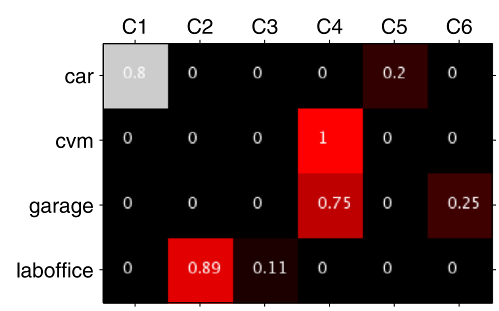

Fig. 7. Co-occurrence matrix obtained considering the *Office Day* scenario with the "New GT".

can observe that these scenes are all related to the same activity (i.e., working at a PC) and the visual content is very similar. We repeated the tests considering an alternative Ground Truth that considers the three above mentioned scene classes to be a unique class (Laboffice). The results of this test are reported in the column labeled as "New GT" of Table 2. In this case the *Office Day* scenario result achieves significant improvements for all the considered evaluation measures. Fig. 7 shows the co-occurrence matrix obtained considered the new Ground Truth.

The last row of Table 2 reports the results obtained by applying the proposed approach on the set of 10 videos obtained by considering only the *Home Day* and the *Office Day* sets of videos. This test have been done to analyse the robustness of the proposed approach considering more contexts and also scenes that appear only in some videos. Furthermore, to better assess the effectiveness of the proposed approach, we performed a number of incremental tests by varying the number of input videos from 2 to 10 videos of the considered set. This correspond to consider from 2 to 10 days of monitoring. The obtained scores reported in Fig. 9 shows that the evaluation measures are quite stable. We also evaluated the performance of the proposed approach by varying the threshold T_σ value (see Fig. 10). Also in this case the results demonstrate that the method is robust. We also tested the between analysis ap-

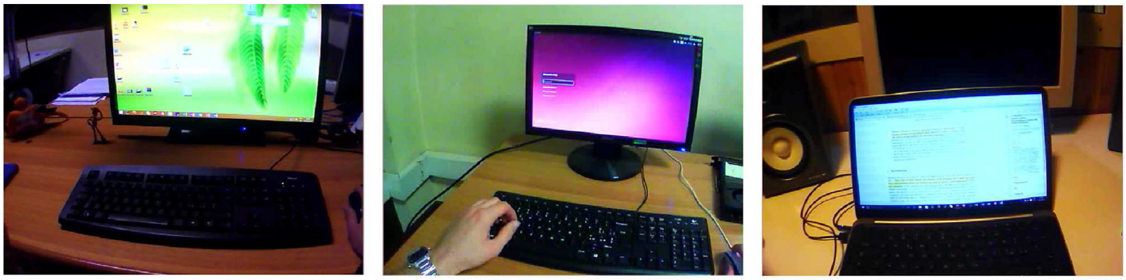

Fig. 8. Some first person view images from the *Office Day* scenario depicting frames belonging to the classes “laboffice”, “office” and “studio”.

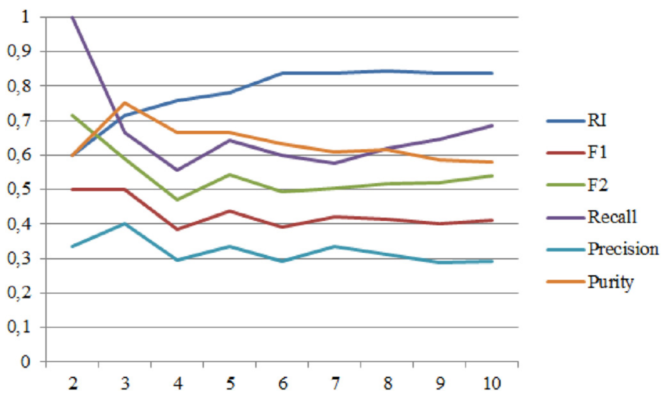

Fig. 9. Evaluation measures at varying of the number of videos.

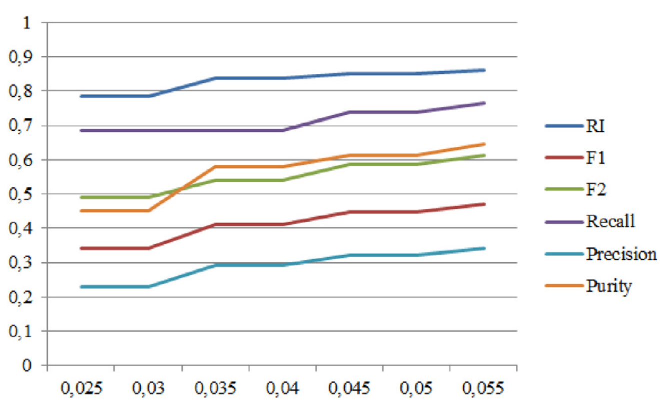

Fig. 10. Evaluation measures obtained by varying the threshold T_σ.

proach by using the color histogram and the distance function used in RECfusion [13]. When the system uses such approach there is a high variance in the obtained distance values even if there isn't any segmentation block to be matched among videos. This causes the matching between uncorrelated blocks and hence errors.

5. Popularity estimation

Considering the improvements achieved by the proposed method in the context of egocentric videos, we have tested the approach to solve the popularity estimation problem defined in [13]. The popularity of the observed scene in an instant of time depends on how many devices are looking at that scene. In the context of organizing the egocentric videos the popularity can be useful to estimate the most popular scene over days.

To properly test our approach for popularity estimation we have considered the dataset introduced in [13]. This allows a fair comparison with the state of the art approaches for this problem be-

Table 3
Experimental results on the popularity estimation problem.

		[13]			Proposed method		
Dataset	dev	Pa/Pr	Pg/Pr	Po/Pr	Pa/Pr	Pg/Pr	Po/Pr
Foosball	4	1.023	1	0.023	**1**	**1**	**0**
Meeting	2	1.011	0.991	0.020	**0.991**	**0.991**	**0**
Meeting	4	**0.988**	**0.955**	0.033	0.930	0.930	**0**
Meeting	5	**0.886**	**0.755**	0.131	0.698	0.704	**0.006**
SAgata	7	1.049	**1**	0.050	**0.989**	0.989	**0**

cause the results can be directly compared with the ones in [13]. Differently than [13], after processing the videos with the intraflow analysis and segmentation refinement proposed in this paper, we have used the proposed CNN features and the clustering approach of [13]. To evaluate the performances of the compared methods we used three measures. For each clustering step we compute:

- P_r: ground truth popularity score (number of cameras looking at the most popular scene) obtained from manual labelling;
- P_a: popularity score computed by the algorithm (number of the elements in the popular cluster);
- P_g: number of correct videos in the popular cluster (inliers).

From the above scores, the weighted mean of the ratios P_a/P_r and P_g/P_r over all the clustering steps are computed [13]. The ratio P_a/P_r provides a score for the popularity estimation, whereas the ratio P_g/P_r assesses the visual content of the videos in the popular cluster. These two scores focus only on the popularity estimation and the number of inliers in the most popular cluster. Since the aim is to infer the popularity of the scenes, it is useful to look also at the number of outliers in the most popular cluster. In fact, the results reported in [13] show that when the algorithm works with a low number of input videos, the most popular cluster is sometimes affected by outliers. These errors could affect the popularity estimation of the clusters and, therefore, the final output. Thus, we introduced a third evaluation measure that takes into account the number of outliers in the most popular cluster. Let P_o be the number of wrong videos in the popular cluster (outliers). From this score, we compute the weighted mean of the ratio P_o/P_r over all the clustering steps, where the weights are given by the length of the segmented blocks (i.e., the weights are computed as suggested in [13]). This value can be considered as a percentage of the presence of outliers in the most popular cluster inferred by the system. The aim is to have this value as lower as possible.

Table 3 shows the results of the popularity estimation obtained by the compared approaches. The first column shows the results obtained by the approach proposed in [13]. Although this method achieves good performance in terms of popularity estimation and inliers rate, the measure P_o/P_r highlights that the method suffers from the presence of outliers in the most popular cluster. This means that the popularity ratio (P_a/P_r) is affected by the presence of some outliers in the most popular cluster. Indeed, in most cases

Fig. 11. Examples of clustering performed by our system and the approach in [13]. Each column shows the input frames taken from different devices. The color of the border of each frame identifies the cluster. The proposed method has correctly identified the three clusters.

the P_a/P_r value is higher than 1 and P_o/P_r is greater than 0. The second column of Table 3 is related to the results obtained with the proposed approach. We obtained values of popularity estimation and inliers rate comparable with respect to [13]. In this case, the values of popularity score are all lower than 1, which means that sometimes the clustering approach lost some inliers.

However it worth to notice that for all the experiments performed with the proposed approach, the outlier ratio P_o/P_r is very close to zero and lower with respect to the values obtained by Ortis et al. [13]. This means that the most popular cluster obtained by the proposed approach is not affected by outliers, and this assure us that the output video belongs to the most popular scene. By performing a visual assessment of the results, we observed that using the proposed CNN features in the clustering phase involves a fine-grained clustering, which better isolates the input videos during the transitions between two scenes or during a noise time interval. This behaviour is not observed in the outputs obtained by [13]. Using the color histogram indeed, the system defines a limited number of clusters that are, therefore, affected by outliers. Some examples about this behaviour are shown in Fig. 11 and Fig. 12. In these figures, each column shows the frames taken from the considered devices at the same instant. The border of each frame identifies the cluster. The first column of each figure shows the result of the proposed approach, and the second column shows the result of the method proposed in [13]. Specifically Fig. 11 shows an example of clustering performed by the compared approaches during the analysis of the scenario *Foosball*. In this example, the first and the third devices are viewing the same scene (the library), the second device is viewing the sofa, whereas the fourth device is performing a transition between the two scenes. In such case, the proposed method creates a different cluster for the device that is performing the transition, whereas the method proposed in [13] includes the scene in the same cluster of the second device. Fig. 12 shows an example of the popularity clustering performed during the analysis of the scenario *Meeting*. In this case both the approaches fail to insert the second frame in the popular cluster due to the huge difference in scale, but the method in [13] creates a cluster that includes the second and the fifth device (despite they are viewing two different scenes) whereas the proposed method can distinguish the second image from the fifth one.

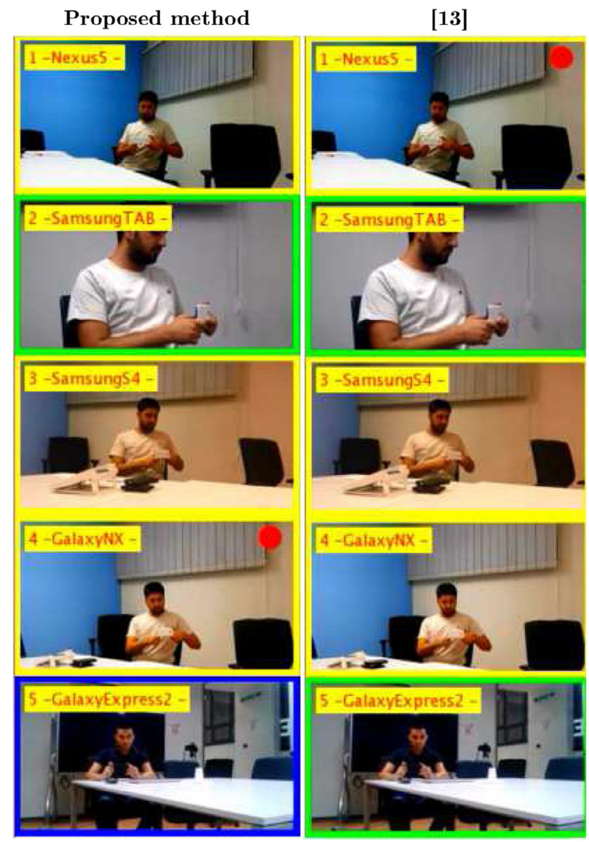

Fig. 12. Examples of clustering performed by our system and the approach in [13]. Each column shows the input frames taken from different devices. The color of the border of each frame identifies the cluster. The proposed method produces better results in terms of outlier detection.

Besides the improvement in terms of performance, proposed approach provides also an outstanding reduction of the computation time. In fact, the time needed to extract a color histogram as suggested in [13] is about 1.5 s, whereas the time needed to extract the CNN feature used in our approach is about 0.08 s. Regarding the popularity intraflow analysis on the mobile video dataset, the proposed approach achieves similar segmentation results compared to the SIFT based approach. However, as in the wearable domain test, the proposed segmentation method strongly outperforms [13] when the system have to deal with noise (e.g., hand tremors).

6. Conclusions and future works

This paper proposes a complete framework to segment and organize a set of egocentric videos of daily living scenes. The experimental results show that the proposed system outperforms the state of the art in terms of segmentation accuracy and computational costs, both on wearable and mobile video datasets. Furthermore, our approach obtains an improvement on outliers rejection on the popularity estimation problem [13]. Future works can consider to extend the system to perform recognition of contexts from egocentric images [2,31–33] and to recognize the activities performed by the user [14].

Appendix A. Pseudocodes

The Algorithm 1 describes the intraflow analysis process (see

Data: Input Video
Result: Segmentation Video
$C \leftarrow newTemplateAt(0)$;
$SEEK \leftarrow True$;
$t \leftarrow 0$;
while $t \leq videoLength - step$ **do**
  $t \leftarrow t + step$;
  $fc7_t \leftarrow extractFeatureAt(t)$;
  if $SEEK$ **then**
    $sim \leftarrow cosSimilarity(C.fc7, fc7_t)$;
    if $sim < T_f$ **then**
      $C \leftarrow newTemplateAt(t)$;
    else
      if *C is a stable template* **then**
        $SEEK \leftarrow False$;
        $TS \leftarrow TS \bigcup \{C\}$;
        $T \leftarrow C$;
        $T.scene\_id \leftarrow sceneAssignment()$;
      end
    end
  else
    $sim \leftarrow cosSimilarity(T.fc7, fc7_t)$;
    $slope \leftarrow slopeAt(t)$;
    if *the slope function has a peak* **then**
      $SEEK \leftarrow True$;
      $C.fc7 \leftarrow fc7_t$;
      $T \leftarrow C$;
    end
    $assignIntervalSceneId(T.scene\_id, t - step, t - 1)$;
  end
end
$applySegmentationRefinements()$;
Algorithm 1: Intraflow Analysis Pseudocode.

Section 2.1) exploiting the involved variables and utility functions as reported in the following:

- A template is represented by a `Template Object`. It keeps information about the time of the related image frame, the *fc7* feature, and the scene ID assigned during the procedure.
- A new `Template Object` is created by the function *newTemplateAt*(t). This function initializes a `Template Object` with the information related to the time t.
- The flag *SEEK* is *True* if the procedure is performing the research of a new stable template. In this case, the `Template Object` `C` represents the current template candidate, which is assigned to the current template instance T if its stability has been verified according to the conditions defined in [13]. In this case, the new template instance T is added to the Templates Set TS. If the flag *SEEK* is *False*, the procedure compares the current reference template T with the forward frames until a peak in the slope sequence is detected.
- The function *sceneAssignment*() performs the scene ID assignment after a new template is defined. In particular, it assigns *Rejected* to all the frames in the interval between the instant when the new template has been requested to the time when it has been defined. Then, the function finds the scene ID of the scene depicted by the new defined template, eventually performing the backward procedure (as described in Section 2.1.3).
- After the segmentation of the input video, the refinements described in Section 2.1.4 are applied by the function *applySegmentationRefinements()*.

The Algorithm 2 describes the between video flows analysis

Data: Set S of Segmented Videos
Result: Clusters of the Segments
initialize all link strengths to zero;
for *each* *video* v_A *in* S **do**
  for *each* *block* b_{v_A} *in* v_A **do**
    if $linkStrength[b_{v_A}]$ *!= 0* **then**
      $scene\_Id \leftarrow b_{v_A}.scene\_Id$;
    else
      assign a new $scene\_Id$ to b_{v_A};
    end
    for *each* *video* v_{B_j} *in* $S \setminus \{v_A\}$ **do**
      $b_{v_{B_j}} = \arg\max_{\bar{b}_{v_{B_j}} \in v_{B_j}} \{CosSimilarity(b_{v_A}, \bar{b}_{v_{B_j}}) \mid \sigma_{(b_{v_A}, v_{B_j})} \geq T_\sigma\}$ **if** $CosSimilarity(b_{v_A}, b_{v_{B_j}}) > linkStrength[b_{v_{B_j}}]$ **then**
        assign $scene\_Id$ to $b_{v_{B_j}}$;
        linkStrength$[b_{v_{B_j}}] \leftarrow CosSimilarity(b_{v_A}, b_{v_{B_j}})$;
      end
    end
  end
end
Algorithm 2: Between Flows Video Analysis pseudocode.

process (see Section 2.2). This procedure takes as input a set S of videos, previously processed with the intraflow analysis described in Algorithm 1. To describe this procedure we defined a data structure called *linkStrength*. This data structure is an array indexed with the blocks extracted from the analysed videos. Considering as example a block segment b_v, the value of $linkStrength[b_v]$ is equal to zero if the block b_v has not yet been assigned to a scene ID. Otherwise, it is equal to the cosine similarity value between b_v and the matched block which caused the assignment. Indeed, all the link strengths are initialized to zero. Then, if the scene ID assigned to b_v changes, the value of $linkStrength[b_v]$ is updated with the new cosine similarity value, as well as the scene ID value assigned to b_v.

References

[1] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, W.W. Mayol-Cuevas, You-do, i-learn: Discovering task relevant objects and their modes of interaction from multi-user egocentric video., in: British Machine Vision Conference, 2014.
[2] A. Furnari, G.M. Farinella, S. Battiato, Recognizing personal contexts from egocentric images, in: Proceedings of the IEEE International Conference on Computer Vision Workshops - Assistive Computer Vision and Robotics, 2015, pp. 393–401.
[3] T. Kanade, Quality of life technology, Proc. IEEE 100 (8) (2012) 2394–2396.
[4] A. Furnari, G.M. Farinella, S. Battiato, Recognizing personal locations from egocentric videos, IEEE Trans. Hum. Mach. Syst. 47 (1) (2017) 6–18.
[5] M.L. Lee, A.K. Dey, Lifelogging memory appliance for people with episodic memory impairment, in: The 10th ACM International Conference on Ubiquitous Computing, 2008, pp. 44–53.
[6] V. Buso, L. Hopper, J. Benois-Pineau, P.-M. Plans, R. Megret, Recognition of activities of daily living in natural “at home” scenario for assessment of alzheimer's disease patients, in: IEEE International Conference on Multimedia & Expo Workshops, IEEE, 2015, pp. 1–6.
[7] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, J.-F. Dartigues, Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia, Multimed. Tools Appl. 69 (3) (2014) 743–771.
[8] Y. Gaëstel, S. Karaman, R. Megret, O.-F. Cherifa, T. Francoise, B.-P. Jenny, J.-F. Dartigues, Autonomy at home and early diagnosis in alzheimer's disease: Utility of video indexing applied to clinical issues, the immed project, in:

Alzheimer's Association International Conference on Alzheimer's Disease 2011, p. S245.
[9] J. Pinquier, S. Karaman, L. Letoupin, P. Guyot, R. Mégret, J. Benois-Pineau, Y. Gaëstel, J.-F. Dartigues, Strategies for multiple feature fusion with hierarchical hmm: application to activity recognition from wearable audiovisual sensors, in: 21st International Conference on Pattern Recognition, IEEE, 2012, pp. 3192–3195.
[10] N. Kapur, E.L. Glisky, B.A. Wilson, External memory aids and computers in memory rehabilitation, in: The Essential Handbook of Memory Disorders for Clinicians, 2004, pp. 301–321.
[11] C. Gurrin, A.F. Smeaton, A.R. Doherty, Lifelogging: personal big data, Found. Trends Inf. Retrieval 8 (1) (2014) 1–125.
[12] M.D. White, Police officer body-worn cameras: assessing the evidence, Washington, DC: Office of Community Oriented Policing Services, 2014.
[13] A. Ortis, G.M. Farinella, V. D'Amico, L. Addesso, G. Torrisi, S. Battiato, Recfusion: automatic video curation driven by visual content popularity, ACM Multimedia, 2015.
[14] A. Fathi, A. Farhadi, J.M. Rehg, Understanding egocentric activities, in: IEEE International Conference on Computer Vision, 2011, pp. 407–414.
[15] H. Pirsiavash, D. Ramanan, Detecting activities of daily living in first-person camera views, in: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on, IEEE, 2012, pp. 2847–2854.
[16] D. Ravi, C. Wong, B. Lo, G.-Z. Yang, Deep learning for human activity recognition: a resource efficient implementation on low-power devices, in: Wearable and Implantable Body Sensor Networks (BSN), 2016 IEEE 13th International Conference on, IEEE, 2016, pp. 71–76.
[17] M.S. Ryoo, L. Matthies, First-person activity recognition: what are they doing to me? in: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2013, pp. 2730–2737.
[18] Y. Poleg, C. Arora, S. Peleg, Temporal segmentation of egocentric videos, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2014.
[19] A. Furnari, G.M. Farinella, S. Battiato, Temporal segmentation of egocentric videos to highlight personal locations of interest, in: International Workshop on Egocentric Perception, Interacion, and Computing - European Conference on Computer Vision Workshop, 2016.
[20] Z. Lu, K. Grauman, Story-driven summarization for egocentric video, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[21] B. Soran, A. Farhadi, L. Shapiro, Generating notifications for missing actions: don't forget to turn the lights off!, in: International Conference on Computer Vision, 2015.
[22] A. Krizhevsky, I. Sutskever, G.E. Hinton, Imagenet classification with deep convolutional neural networks, in: Advances in Naural Information Processing Systems, 2012, pp. 1097–1105.
[23] L. Deng, A tutorial survey of architectures, algorithms, and applications for deep learning, APSIPA Trans. Signal Inf. Process. 3 (2014) e2.
[24] P. Agrawal, R. Girshick, J. Malik, Analyzing the performance of multilayer neural networks for object recognition, in: European Conference on Computer Vision, 2014, pp. 329–344.
[25] S. Bell, P. Upchurch, N. Snavely, K. Bala, Material recognition in the wild with the materials in context database, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[26] J. Hosang, M. Omran, R. Benenson, B. Schiele, Taking a deeper look at pedestrians, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[27] J. Long, E. Shelhamer, T. Darrell, Fully convolutional networks for semantic segmentation, in: The IEEE Conference on Computer Vision and Pattern Recognition, 2015.
[28] J. Yosinski, J. Clune, Y. Bengio, H. Lipson, How transferable are features in deep neural networks? in: Advances in Neural Information Processing Systems, 2014, pp. 3320–3328.
[29] K. Simonyan, A. Zisserman, Very deep convolutional networks for large-scale image recognition, ArXiv:1409.1556 2014.
[30] C.D. Manning, P. Raghavan, H. Schütze, et al., Introduction to information retrieval, Cambridge University Press Cambridge, 2008.
[31] G.M. Farinella, S. Battiato, Scene classification in compressed and constrained domain, IET Comput. Vision 5 (5) (2011) 320–334.
[32] G.M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, S. Battiato, Representing scenes for real-time context classification on mobile devices, Pattern Recognit. 48 (4) (2015) 1086–1100.
[33] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, A. Oliva, Learning deep features for scene recognition using places database, in: Advances in Neural Information Processing Systems, 2014, pp. 487–495.

Alessandro Ortis is a first year Ph.D Student in Computer Science at the University of Catania. He has been working in the field of Computer Vision research since 2012, when he joined to the IPLab (Image Processing Laboratory). He obtained the Master Degree in Computer Science (summa cum laude) from the University of Catania in March 2015. Alessandro was awarded with the *Archimede Prize* for the excellence of academic career conferred by the University of Catania in 2015. Along the way he has done two research internships at *STMicroelectronics* in 2011/2012 and at *Telecom Italia* in 2015. So far he has attended four editions of the ICVSS (*International Computer Vision Summer School*) and one edition of the MISS (*Medical Imaging Summer School*). His research interests lie in the fields of Computer Vision, Machine Learning and Crowdsourced Media Analysis. He is co-author of 5 papers published in international conferences and co-inventor 1 International Patent.

Giovanni Maria Farinella is Assistant Professor at the Department of Mathematics and Computer Science, University of Catania, Italy. He received the (egregia cum laude) Master of Science degree in Computer Science from the University of Catania in April 2004. He was awarded the Ph.D. in Computer Science from the University of Catania in October 2008. From 2008 he serves as Professor of Computer Science for undergraduate courses at the University of Catania. He is also an Adjunct Professor at the School of the Art of Catania in the field of Computer Vision for Artists and Designers (Since 2004). From 2007 he is a research member of the Joint Laboratory STMicroelectronics - University of Catania, Italy. His research interests lie in the field of Computer Vision, Pattern Recognition and Machine Learning. He is author of one book (monograph), editor of 5 international volumes, editor of 5 international journals, author or co-author of more than 100 papers in international book chapters, international journals and international conference proceedings, and of 18 papers in national book chapters, national journals and national conference proceedings. He is co-inventor of 4 patents involving industrial partners. Dr. Farinella serves as a reviewer and on the board programme committee for major international journals and international conferences (CVPR, ICCV, ECCV, BMVC). He has been Video Proceedings Chair for the International Conferences ECCV 2012 and ACM MM 2013, General Chair of the International Workshop on Assistive Computer Vision and Robotics (ACVR - held in conjunction with ECCV 2014, ICCV 2015 and ECCV 2016), and chair of the International Workshop on Multimedia Assisted Dietary Management (MADiMa) 2015/2016. He has been Speaker at international events, as well as invited lecturer at industrial institutions. Giovanni Maria Farinella founded (in 2006) and currently directs the International Computer Vision Summer School (ICVSS). He also founded (in 2014) and currently directs the Medical Imaging Summer School (MISS). Dr. Farinella is an IEEE Senior Member and a CVF/IAPR/GIRPR/AIxIA/BMVA member.

Valeria D'Amico is the head of the Telecom Italia Joint Open Lab. Her interest include Smart Cities, Corporate Entrepreneurship e Social Innovation. She has a Degree in Electronic Engineering at the University of Catania, and she has a Master in Business Administration (MBA) from SDA Bocconi School of Management.

Luca Addesso is a Telecom Italia researcher at the Joint Open Lab WAVE laboratory of Catania. He received his degree in Electronic Engineering from the University of Florence and is co-author of 4 conference proceedings. His research interests lie in the fields of Wearable Devices, Gesture Recognition, Mobile Multimedia and Fast Prototyping.

Giovanni Torrisi is a Telecom Italia researcher at the Joint Open Lab WAVE laboratory of Catania. He received his degree in Computer Science Engineering (summa cum laude) from the University of Catania and is co-author of 3 conference proceedings. His research interests lie in the fields of Wearable Devices, Mobile Design and Development, Data Visualization.

Sebastiano Battiato is Full Professor of Computer Science at University of Catania. He received his degree in computer science (summa cum laude) in 1995 from University of Catania and his Ph.D. in Computer Science and Applied Mathematics from University of Naples in 1999. From 1999 to 2003 he was the leader of the "Imaging" team at STMicroelectronics in Catania. He joined the Department of Mathematics and Computer Science at the University of Catania in 2004 (respectively as assistant professor, associate professor in 2011 and full professor in 2016). He is currently Chairman of the undergraduate program in Computer Science, and Rector's delegate for Education (postgraduates and Phd). He is involved in research and directorship of the IPLab research lab (http://iplab.dmi.unict.it). He coordinates IPLab participation to large scale projects funded by national and international funding bodies, as well as by private companies. Prof. Battiato has participated as principal investigator in many international and national research projects.His research interests include image enhancement and processing, image coding, camera imaging technology and multimedia forensics. He has edited 6 books and co-authored about 200 papers in international journals, conference proceedings and book chapters. Guest editor of several special issues published on International Journals. He is also co-inventor of 22 international patents, reviewer for several international journals, and he has been regularly a member of numerous international conference committees. Chair of several international events (ICIAP 2017, VINEPA 2016, ACIVS 2015, VAAM2014-2015-2016, VISAPP2012- 2015, IWCV2012, ECCV2012, ICIAP 2011, ACM MiFor 2010–2011, SPIE EI Digital Photography 2011- 2012–2013, etc.). He is an associate editor of the SPIE Journal of Electronic Imaging. He is the recipient of the 2011 Best Associate Editor Award of the IEEE Transactions on Circuits and Systems for Video Technology. He is director (and co-founder) of the International Computer Vision Summer School (ICVSS), Sicily, Italy. He is a senior member of the IEEE.