Organizing Egocentric Videos for Daily Living Monitoring

Alessandro Ortis
Universitá degli Studi di Catania
Viale A. Doria 6, 95125
Catania, Italy
ortis@dmi.unict.it

G.M. Farinella
Universitá degli Studi di Catania
Viale A. Doria 6, 95125
Catania, Italy
gfarinella@dmi.unict.it

Valeria D'Amico
Telecom Italia
JOL WAVE - Viale A. Doria 6, 95125
Catania, Italy
valeria1.damico@telecomitalia.it

Luca Addesso
Telecom Italia
JOL WAVE - Viale A. Doria 6, 95125
Catania, Italy
luca.addesso@telecomitalia.it

Giovanni Torrisi
Telecom Italia
JOL WAVE - Viale A. Doria 6, 95125
Catania, Italy
giovanni.torrisi@telecomitalia.it

Sebastiano Battiato
Universitá degli Studi di Catania
Viale A. Doria 6, 95125
Catania, Italy
battiato@dmi.unict.it

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 assistive applications in this context there is the daily living monitoring of a user that is wearing the camera. In this paper we propose a system devoted to automatically organize videos acquired by the user over different days. By employing an unsupervised segmentation, each egocentric video is divided in chapters by considering the visual content. The video segments related to the different days are hence linked to produce graphs which are coherent with respect to the context in which the user acts. Experiments on two different datasets demonstrate the effectiveness of the proposed approach which outperforms the state of the art, both in accuracy and computational time with a good margin.

Keywords

Social Cameras, Video Curation, Summarization, First Person Vision, Assistive Computer Vision

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 (e.g., smart glasses) together with an increasing diversity of such devices with respect to hardware capabilities and computational resources. 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 [4] [11] [17]. 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:

- 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 perpective.
- Environment/Objects Aware: systems able to understand what objects sourround the user, where they are with respect the user's perspective and what the environment looks like.
- Target Aware: systems able to understand what others are doing, and how they interact with the user that is wearing the device.

Wearable cameras provide a practical method to collect first person view video datasets related to people during their daily living in different environments. The growth of wearable cameras leads to a number of challenging tasks for the research community, as well as a huge number of applications. The monitoring of the activities and the events that a person experiences, can help to stimulate the memory of users that suffer from memory disorders [21]. Several works on recognition and indexing of daily living activities of patients with dementia have been recently proposed [3] [19] [13] [27]. The exploitation of aids for people with memory problems is proved to be one of the most effective ways to aid rehabilitation. It has been demonstrated to be effective in increasing independence in brain injured patients [18]. Furthermore, by recording and organizing the daily habits performed by a patient, a doctor can 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 by a patient can be analysed by experts to monitor the user's daily living for assistive purposes. Live recording for life logging applications poses challenges on how to perform automatic index and summarization of these big personal multimedia data [14]. Indeed, the task of monitoring daily living with first person cameras involves several steps to be performed, such as the

LTA'16, October 16 2016, Amsterdam, Netherlands

DOI: http://dx.doi.org/10.1145/2983576.2983578

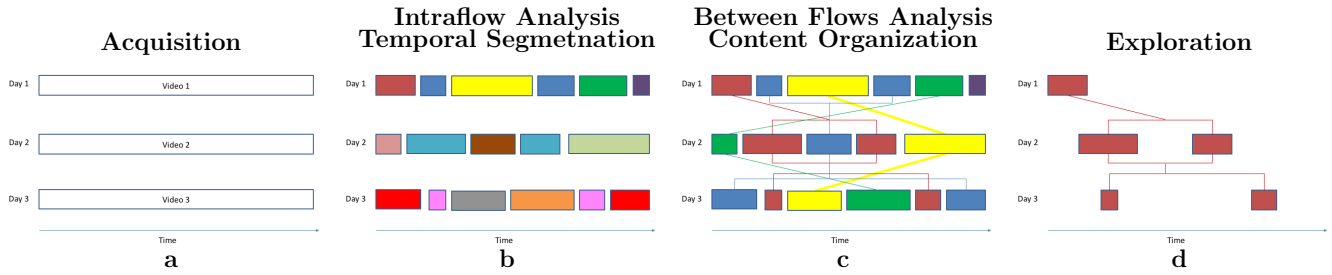

Figure 1: Overall scheme of the proposed framework. Given a set of first person videos recorded among different days (a), the system performs a segmentation of each video according to the scenes visual context (b). Then, the different segments are organized by matching their visual contents among different days (c). As result, the system provides an organization for the monitoring of the user's daily living that can be explored in a user interface (d).

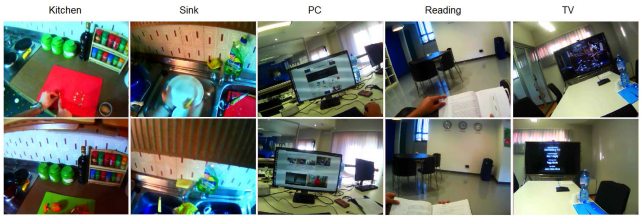

Figure 2: Some first person view images from the considered dataset. Each column reports two different shots of the same scene acquired among different days.

segmentation of the videos in chapters, indexing of frames and summarization of long videos.

In the last years, due to the growing diffusion of wearable cameras and the expansion of their application areas, several papers have been published by considering different vision tasks from first person perspective. The work in [28] proposes a temporal segmentation method of egocentric videos by analysing the motion of the camera weared by a user. The authors suggest a method that segments the input egocentric video with respect to 12 different activities organized hiararchically upon cues based on wearer's motion (e.g., static, sitting, standing, walking, etc.). In [24] the authors propose a method that takes a long input video and returns a short video summary as output by selecting a set of video subshots depicting the essential moments. The final output is obtained by defining a coherent chain of video subshots in which each subshot influences the next through some subset of influential visual objects. The method proposed in [31] learns the sequences of actions involved in a set of habitual daily activities. Thus, it is able to recognize the current action performed by the wearer, predict the next action and generate a notification if there is a missing action in the sequence. The framework presented in [26] (RECfusion), and further extended in [25], is able to automatically process multiple video flows from different devices to understand the most popular scenes for a group of end-users. The aim of RECfusion is to catch the audience's interest during a social event by exploiting a set of crowdsourced video flows. Its output is a video which represents the most popular scenes organized over time. This method has been successful applied on videos acquired with mobile cameras, but the performances decrease when videos are acquired with wearable cameras.

In this paper we build on the RECfusion methodology [26] improving it in the context of daily living monitoring from egocentric videos. In [26] the multiple videos are analysed by using two algorithms: the former is used to segment the different scenes, transitions and the unstable intervals within each video (intraflow analysis). The latter is employed to perform the grouping of the videos related to the involved devices over time by taking into account the visual content of the previously segmented video streams. The popularity of the obtained clusters over time is used to produce the final video. During the segmentation phase, each scene is represented by a set of extracted SIFT descriptors [23] which is considered as a template of the corresponding scenes within the segment. The intraflow analysis consists in the comparison of templates extracted from different frames of a video to split it in blocks with frames having similar visual content. As reported in the experimental results of the original paper [26], the intraflow analysis of RECfusion suffers when applied to egocentric videos because they are highly unstable due to the user's movements.

To compare different blocks of different devices (between flow analysis), the authors of [26] exploit an image representation based on a quantized weighted color histogram. The weights are obtained by using a gradient map computed in both x and y directions as proposed in [6]. This representation is extracted after an equalization phase proposed in [10]. Such an image representation requires high computational effort to be extracted, in fact it involves several processes: image equalization, gradient extraction, weighted histogram computation and quantization. Furthermore, the computational effort augments drastically with the dimensions of the input images.

In this paper we propose a framework which overcomes the problems of [26]. The overall pipeline is shown in Figure 1. 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 with the same contents can be then visualized by exploiting an interactive web-based user interface. Experiments have been performed on egocentric videos acquired in two different scenarios (Figure 2).

Differently than [26], the proposed framework allows to have better performances for egocentric videos organization. We obtained an improvement with respect to RECfusion on both segmentation accuracy and computational costs. This is obtained by using a unique representation for frames of

the videos based on CNN features [20] (for both intraflow and between flows analysis) instead of two different representations based on SIFT and the color histogram. Moreover, experiments show that the proposed framework outperforms RECfusion [26] also for mobile video organization.

The rest of the paper is organized as follows. Section 2 presents the proposed framework. In particular, Section 2.1 and Section 2.2 present the proposed segmentation and content organization approaches respectively. Section 3 introduces the considered wearable dataset, whereas the discussion of the obtained results is given in Section 4. In Section 5 the proposed system is further compared with respect to RECfusion [26] on mobile videos. Finally Section 6 concludes the paper and gives insights for further works.

2. PROPOSED FRAMEWORK

The proposed framework employs two main steps on the videos acquired by a wearable camera: temporal segmentation and content organization. Figure 1 shows the scheme of the overall pipeline of the proposed system. Starting from a set of egocentric videos recorded among multiple days (Figure 1 (a)), the first step performs an intraflow analysis of each video to segment them with respect to the different scenes observed by the user (Figure 1 (b)). Then, the obtained segments over different days that regard to the same content are grouped (Figure 1 (c)).The system 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 an efficient browsing of the segmented videos by exploiting the inferred organization (Figure 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 segmentation of each input video by associating a scene ID to each video segment. This means that the identification of the segments given by the intraflow analysis doesn't associate video segments extracted from two different video flows. Instead, it works to segment each video and to associate segments with same content within each video. In the following we focus our analysis on the issues related to the intraflow algorithm proposed in [26] when applied on first person videos. This is useful to introduce to the reader on the main problems of a classic feature based matching approach for temporal segmentation in wearable domain. Then we present the proposed solution for the intraflow analysis to split each video in blocks according to their visual content. The main differences between the approach in [26] and the proposed one are related to both, the employed image representations and the way we obtain the final segments of each video.

2.1.1 Issues related to SIFT based templates

The intraflow analysis proposed in [26] compares two scenes considering the number of matchings between a reference template (i.e., the last stable set of SIFT descriptors extracted from the last detected scene) and the current frame.

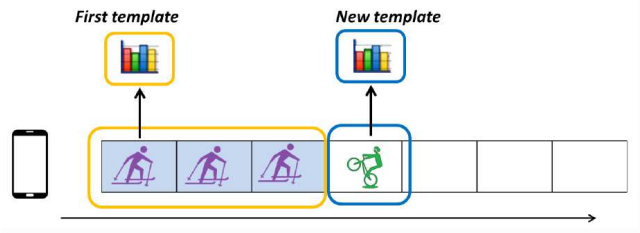

Figure 3: Template based video segmentation performed by the intraflow analysis.

When the algorithm detects a sudden decrease in the number of matchings, it refreshes the reference template extracting a new set of SIFTs and cuts the video (see Figure 3). 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. The scene template is a set of SIFT descriptors that must accomplish specific properties of reliability [26]. When a new template is defined, it is compared with the past templates in order to understand if it regards a new scene or a known one (backward search phase). In this way the IDs are assigned to the different segments.

Although this method works very well with videos acquired with mobile cameras, it has difficulties when applied on videos acquired with wearable cameras. 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 found during the backward search of the template performed by [26]. Furthermore, the approach in [26] requires to perform several SIFT descriptor extraction and matching operations (including geometric verifications) to exclude false positive matchings [26]. An example of a segmentation obtained with the algorithm in [26] on a egocentric video is reported in Figure 4. The first row shows the Ground Truth segmentation of the video acquired with a wearable camera in an home environment[1]. The second row shows the segmentation result of the SIFT based interflow analysis proposed in [26]. The algorithm works well when the scene is quite stable (e.g., when the user is cooking in the kitchen, or when he is sit on the sofa 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 Figure 4 (which correspond to a user cleaning dishes at the sink) the user is continuously moving his head.

Furthermore, in the aforementioned example the intraflow approach based on SIFT features detects a total of 8 different scenes instead of 3. In such cases, the algorithm can't find the matchings between the current frame and the reference template due to two main reasons:

- When the video is unstable, even though the scene doesn't change, the matchings between local features

[1]The video related to the example in Figure 4 is available for visual inspection at the URL http://iplab.dmi.unict.it/recfextension.

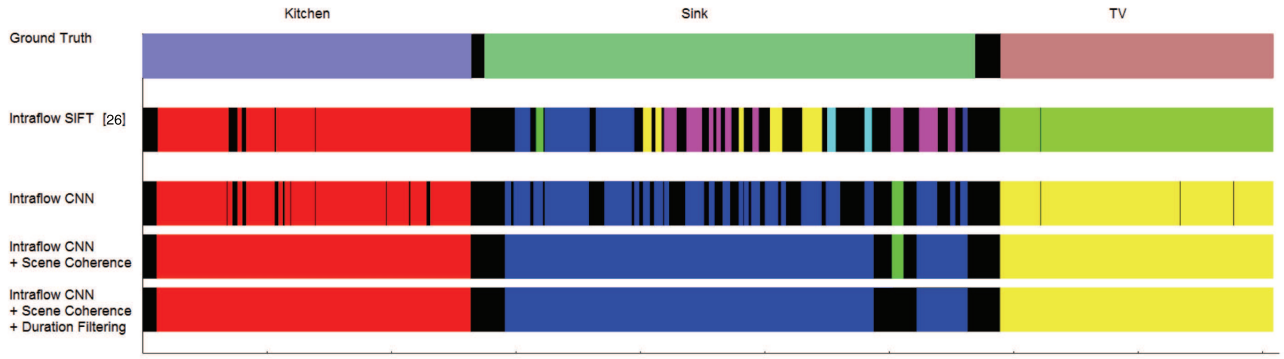

Figure 4: Output of the intraflow analysis using CNN and SIFT applied to the video *Home Day 1*. The first row is the Ground Truth of the video, which depicts ten minutes of tipical home activities (cooking in the kitchen, cleaning dishes at the sink, watching a TV program in the living room). The second row shows the result of the intraflow analysis with the method proposed in [26]. The third row shows the result of the intraflow analysis performed by the proposed system, whereas the last two rows show the results of the application of the scene coherence and duration filtering criteria. The performances of the intraflow methods applied to this video instance, are detailed in the first row of Table 1.

are not reliable and stable along time;

- In a closed space such as an indoor environment, the different objects of the scene can be very close to the viewer. Thus, just 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 2.1.1, we exploit an holistic feature to describe the whole image rather than an approach based on local features. In particular, for the intraflow analysis we represent frames by using features extracted with a Convolutional Neural Network [5]. Over the last few years, the increasing computational power of GPUs and the creation of large image datasets have allowed Convolutional Neural Networks (CNNs) to show outstanding performance in many Computer Vision challenges. Furthermore, CNNs have proved to be very effective for transfer learning problems. In this work we consider the CNN proposed in [20], which is called *AlexNet*. It consists of seven internal layers with a final 1000-way softmax which produces a distribution over the 1000 predefined classes of the ImageNet dataset [29]. In our experiments, we exploit the representation obtained from the last hidden layer of *AlexNet*, which consists of a 4096 dimensional feature vector (*fc7* feature).

We decided to use *AlexNet* representation for the following motivations:

- The *fc7* representation of *AlexNet* has been successfully used as a general image representation for classification purpose in the last few years [1] [2]. Thus, the use of this feature is highly known and tested by the community.
- The feature extracted by *AlexNet* have been used successfully for transfer learning [15] [22] [32].
- *AlexNet* architecture is a short network compared to other networks (e.g., VGG [30]). Thus, it allows to perform the feature extraction very quickly. Considering that the proposed system needs to extract a huge number of image representations from the frames of several video flows, computational costs and time are critical factors.

In [26] the similarity between the scene templates is defined by the number of matchings between two SIFT sets, computed after a geometrical verification check. The proposed solution, instead, compares a pair of *fc7* features by using 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 two *fc7* 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 following frames. When the algorithm detects a sudden decrease in the cosine similarity sequence, it refreshes the reference template selecting a new stable *fc7* feature. As in the approach presented in [26], in order to detect such changes the system computes the value of the slope (i.e., the variation of the cosine similarity in a range interval). According to [26], when the slope has a positive value (which correspond to a sudden decrease of the cosine similarity) the algorithm finds a new template. 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.

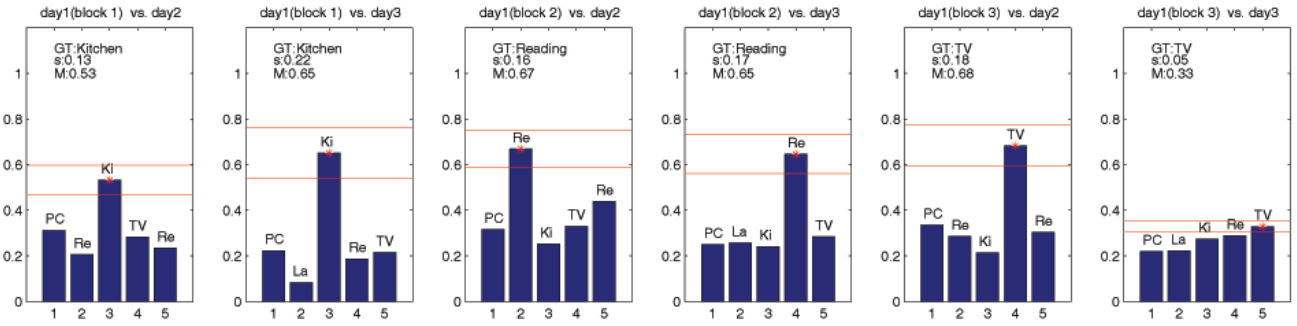

Figure 5: Linking segments between egocentric videos. Each plot shows the cosine similarity values obtained by comparing one block of a video v_A with respect to some blocks extracted from the other videos. For each plot the following details are specified: the Ground Truth of the considered block (GT), the value of $\sigma_{(b_{v_A}, v_{B_j})}$ (s), the value of the maximum similarity (M). The red asterisks indicate the segment blocks that are selected by considering the equation (2). The red lines depict the boundings of a neighborhood of the maximum value with a range of $\pm\sigma_{(b_{v_A}, v_{B_j})}/2$. The following abbreviations are used: Ki - Kitchen, Re - Reading, La - Laboratory.

2. A template is compared with the past templates during the backward checking phase.

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. When the algorithm compares a template with the forward frames, it 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 with respect to the approach proposed in [26].

- In [26] 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 frames between frames caused by the movement of the head of the user (Noise) or by 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 this rejected group are considered as "Transition" if the duration of the block is longer than 3 seconds (i.e., we consider that head movements much faster than 3 seconds). Otherwise they are classified as "Noise". In case of noise, the algorithm doesn't assign a new scene ID to the frame that follows the "Noise" video segment because the noise is related to head movements but the user is in the same context. This simple heuristic is very useful when the system considers videos acquired with wearable devices. On the contrary, when a video has been acquired by using a mobile camera, the viewed scene can be quickly changed by the movement of the hand. Since we have to deal with first-person view videos, when the user changes its location he changes his position in the environment. Thus, the transition between different scenes involves a longer time interval, and this is the reason of the established distintion between noise and transition with the proposed heuristic.

- The other introduced rule is related to the backward verification. In [26] it is performed starting from the last found template and proceding backward. It 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 [26] because it compares sets of local features and relyes on the number of achieved matchings. The approach proposed in this paper compares pairs of vectors instead of a number of descriptors, and selects the past template that yelds the best similarity to the new one. In particular, the proposed approach compares the new template with all the previous ones, and all the past templates that yelds a cosine similarity grather than T_b are considered. 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. This make more robust the ID assignment in the intraflow analysis.

Considering the example in Figure 4, the segmentation results obtained with the proposed intraflow approach (third row) are much better than the one obtained using SIFT features (second row). After the intraflow analysis segmentation a refinement is performed as detailed in the following subsection.

2.1.4 *Segmentation Refinement*

The previous section describes the proposed method to perform the intraflow analysis in order to obtain segments with coherent visual content. Starting from the result of the intraflow analysis (see the third row of Figure 4), 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

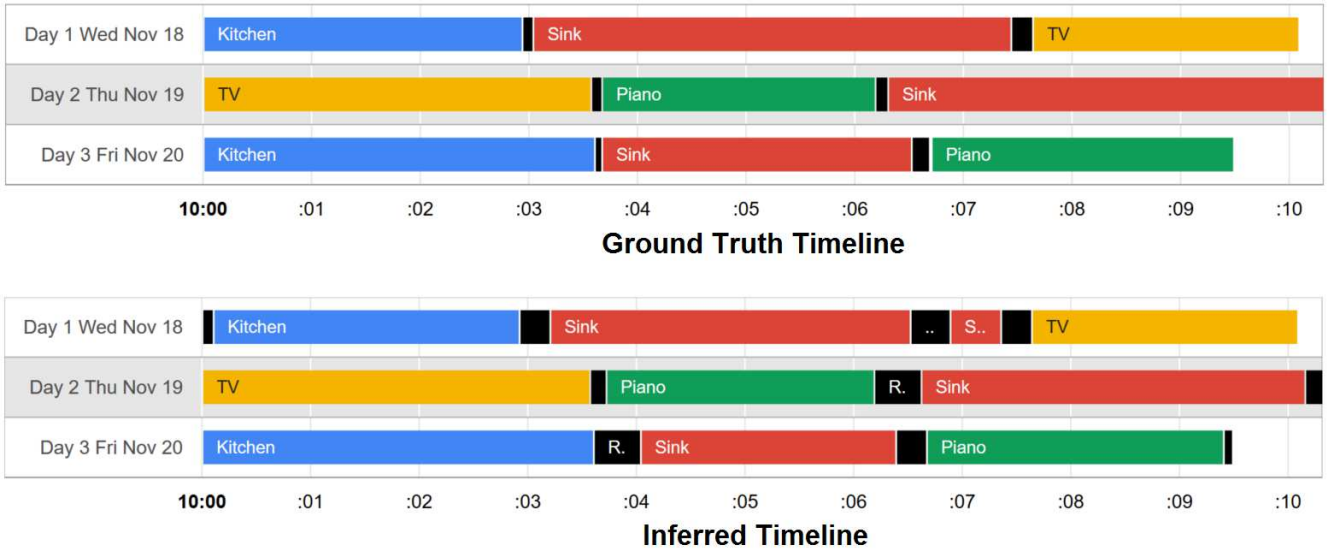

Figure 6: The top timelines show the Ground Truth segmentation and organization related to the *Home Day* scenario, whereas the bottom timelines show the inferred segmentation and organization obtained by the proposed method.

is marked as a "Transition block". We refer to this criteria as *Scene Coherence*. The result of this step in the example considered in Figure 4 is shown in the fourth row. When comparing the segmentation with the Ground Truth, the improvement with respect to [26] (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 seconds (*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 significative duration in order to be able to observe the behaviour. This step follows the *Scene Coherence* step. The final result is shown in the last row of Figure 4. Despite some frames are incorrectly rejected (during scene changes) the proposed pipeline is much more robust than [26]. This is quantitatively demonstrated on two different datasets in the experimental section of this paper. Comparing the final result with respect to the Ground Truth (in Figure 4), we can observe that the proposed system correctly segments almost all the three scenes of the example video.

2.2 Between Video Analysis

When each egocentric video is segmented the next challenge is to determine which segments among the different days (i.e.,videos) 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 (i.e., the reference template that has been used to represent that scene for the longer time interval during the slope checking with the cosine similarity sequence, explained in 2.1.3). 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}:

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

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}. 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 is repeated until all the videos in the pool are processed. After this process, a scene ID is assigned to all the blocks of all the considered videos. This identification is unique for the set of egocentric videos. Thus, a pair of blocks with the same ID are linked, even if they belong to different videos, and all the segments are connected in a graph with multiple components (as in Figure 1 (c)). Figure 5 shows the details of the comparisons performed with the above approach. 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 isn't matched (e.g., the first plot of the first row, and the first two plots of the second row), all the similarity values of the missmatched blocks are similar. This causes low values of σ and helps the system to understand that the searched activity is not present. When the similarity is low, the blocks are skipped. In our experiments, we used $T_\sigma = 4 \times 10^{-2}$. However we observed that all values of T_σ between 3×10^{-2} and 5×10^{-2} also provide similar per-

formances. The use of such a threshold on the σ value generalizes the obtained measures of similarity between blocks. Indeed, the system considers two blocks to be similar if their similarity has a high value with respect to the set of similarity values obtained comparing different pairs of blocks. Furthermore, this threshold is needed to detect isolated scenes (i.e., scenes with only one instance among all the considered videos). When the searched scene is matched we usually observe high values of σ (higher than T_σ by one order or more), but without a threshold value for σ we couldn't find any isolated block (unless we use a threshold on the similarity value).

3. DATASET

To demonstrate the effectiveness of the proposed approach on first person videos, we have acquired egocentric videos to perform experiments on two different scenarios. All videos have been taken in different days using the same head mounted camera. Specifically, we used a Looxcie LX2 with a resolution of 640x480 pixels. The duration of each video is about 10 minutes.

We considered the following scenarios:

- Home Day: a set of three egocentric videos taken in an home environment. In this scenario, the user performs typical home activities such as cooking, cleaning dishes, watching TV, and playing piano. This set of videos has been borrowed from the dataset used in [12] which is available at the following URL: http://iplab.dmi.unict.it/PersonalContexts/segmentation/.
- Working Day: a set of three videos taken in an office environment. Also in this scenario, the user performs different activities such as reading a book, working in a laboratory, sitting in front of a computer, etc.

Figure 2 shows some examples of the acquired scenes among different days. Each video has been manually segmented to define the blocks that have 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. This Ground Truth is used to evaluate the performances of the framework. The used egocentric videos, as well as the Ground Truth, are available at the following URL: http://iplab.dmi.unict.it/recfextension.

4. EXPERIMENTAL RESULTS

In this section we report the results obtained with the proposed segmentation and between analysis on the aforementioned dataset.

4.1 Segmentation Results

Table 1 shows the performances of the proposed segmentation method. We compared our solution with respect to the method presented in [26]. For each method we computed the quality of the segmentation as the percentage of the well classified frames (Q), the number of detected scenes (S) and the computation time (T). In comparison to [26], we observe strong improvements up to over 16% for segmentation quality (results at fifth row in Table 1) obtained by just applying the segmentation approach explained in section 2.1.3. Furthermore, the application of the segmentation refinements provides improvements up to 30% in segmentation quality (results at third row in Table 1). Considering the mean performaces (last row in Table 1) our system achieves an improvement of over 9% points without segmentation refinements, with over 16% of margin after the segmentation refinements. The proposed method also achieves up to over than 19 minutes in computation time saving (first row in Table 1) and there is an average computation time saving of about 17 minutes with respect to [26].

Furthermore, Table 1 shows that the application of the *Scene Coherence* and the *Duration Filtering* criteria used in the segmentation refinement step allows to detect the correct number of scenes (S). Thus, the application of such simple heuristics, allows to define the exact number of periods of time related to the user's activities, avoiding to select worthless blocks. In sum, as shown in the segmentation example in Figure 4, and as demosntrated by the results reported in Table 1, the proposed system is robust for the segmentation of egocentric videos, and provides high performances with low computational time with respect to [26].

4.2 Between Video Analysis Results

In our experiments, all the segmented blocks among the considered days have been correctly matched by the proposed approach for both scenarios. This means that all the segments have been correctly linked among the different days. Figure 6 shows two timelines related to the videos of the scenario *Home Day*. The first timeline 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 rejected by the algorithm. In order to better asses the results obtained by the proposed system, the reader can perform a visual inspection of the videos produced by our approach at the following URL:
http://iplab.dmi.unict.it/recfextension.
Through the web interface the different segments can be explored. We also tested the between analysis approach by using the color histogram and the distance function used in [26]. 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 on both intraflow and between flow analysis in the context of egocentric videos, we have tested the approach to solve the popularity estimation problem defined in [26]. Specifically we have used the proposed framework to infer the popularity of the scenes. The popularity of the scene in a instant of time depends on how many people are looking at that scene, and therefore can be obtained though the "visual consensus" among multiple video streams acquired by different mobile devices [26]. The proposed approach has been tested for popularity estimation considering the dataset introduced in [26] in order to perform a fair comparison with the state of the art approaches for this problem [16] [26]. We have used the clustering approach of [26]. Differently than [26] we have used the proposed CNN features for clustering after processing the videos with the intraflow analysis proposed in this paper.

Table 1: Intraflow performances for wearable devices, computed using [26] and the proposed approach. Each test is evaluated considering the accuracy of the segmentation (Q), the computation time (T) and the number of the scenes detected by the algorithm (S). The accuracy is measured as the percentage of well classified frames with respect to the Ground Truth. The measured time includes the feature extraction process.

		Intraflow proposed in [26]			Proposed Intraflow Approach		Proposed Interflow Approach with Segmentation Refinement		
Video	**Scenes**	**Q**	**S**	**T**	**Q**	**S**	**Q**	**S**	**T**
HomeDay1	3	62,5%	8	20'45"	77,5%	4	**92,5%**	**3**	**1'23"**
HomeDay2	3	71,6%	3	20'18"	80,3%	4	**94,5%**	**3**	**1'46"**
HomeDay3	3	64,3%	5	19'03"	79,7%	5	**94,3%**	**3**	**1'21"**
WorkingDay1	4	95,7%	5	16'16'	98,4%	5	**99,5%**	**4**	**1'22"**
WorkingDay2	4	82,5%	5	15'15"	98,9%	5	**100%**	**4**	**1'08"**
WorkingDay3	5	98,7%	6	19'02"	99,2%	6	**99,4%**	**5**	**1'29"**
Average		79,2%		18'27"	89,0%		**96,7%**		**1'25"**

Table 2: Experimental results on the popularity estimation problem defined in [26]. The proposed approach is compared with respect to the popularity estimation method in [26].

		[26]			Proposed method		
Dataset	**dev**	**Pa/Pr**	**Pg/Pr**	**Po/Pr**	**Pa/Pr**	**Pg/Pr**	**Po/Pr**
Foosball	4	1,02	1	0,023	**1**	**1**	**0**
Meeting	2	1,01	0,99	0,020	**0,99**	**0,99**	**0,018**
Meeting	4	**0,99**	**0,95**	0,033	0,93	0,93	**0**
Meeting	5	**0,89**	**0,76**	0,131	0,70	0,70	**0,006**
SAgata	7	1,05	**1**	0,050	**0,99**	0,99	**0**

To evaluate the performances of the compared methods, we compute three measures. Two of them are measures proposed in [26]. Specifically, 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 in [26]. Thus, the ratio P_a/P_r provides a score for the popularty estimation, whereas the ratio P_g/P_r assesses the visual content of the videos in the popular cluster. These two scores only focus 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 [26] 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 [26]). 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 2 shows the results of the popularity estimation obtained by considering the proposed approach in comparison with respect to [26]. For each test we computed the three aforementioned performance measures. The first column shows the results obtained by the approach proposed in [26], which performs a SIFT based intraflow analysis and exploits a color histogram representation during the clustering phase. Although the method in [26] achieves good performance in terms of popularity estimation and inliers rate, the measure P_o/P_r demonstrate that it suffers from the prencence 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 the P_a/P_r value is higher than 1 and P_o/P_r is greater than 0.

The second column is related to the results obtained with the proposed approach. We obtained values of popularity estimation and inliers rate comparable with [26]. In this case, the values of popularity score are all lower than 1, this means that sometimes the clustering approach lost some inliers. But it worth to notice that for all the experiments the outlier ratio P_o/P_r is very close to zero and lower that the values obtained by [26]. 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 [26]. 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 Figure 7 and Figure 8. 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 [26]. Specifically Figure 7 shows an example of clustering performed by the compared approaches during the analysis of the scenario *Foosball* [26]. 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 [26] includes the scene in the same cluster of the second device. Figure 8 shows an example of the popularity clustering performed during the analysis of the scenario *Meeting* [26]. In this case, the first four devices are viewing the same scene, but all the approaches fail to insert the second frame in the popular cluster due to the huge difference in scale. This example shows that the method in [26] 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. This demonstrates that the CNN features have a more discriminative power than the color histograms used in [26].

Besides the improvement in terms of performance, the exploitation of the CNN representation provides also an outstanding reduction of the computation time. In fact, the time needed to extract a color histogram as suggested in [26] is about 1.5 seconds, whereas the time needed to extract the CNN feature is about 0.08 seconds. Regarding the popularity estimation experiment, the proposed intraflow analysis achieves similar segmentation results compared to the SIFT based approach [26]. However, as in the wearable domain, the proposed segmentation method strongly outperforms [26] when the system have to deal with noise (e.g., tremors).

Figure 7: Examples of clustering performed by our system and the approach in [26]. 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.

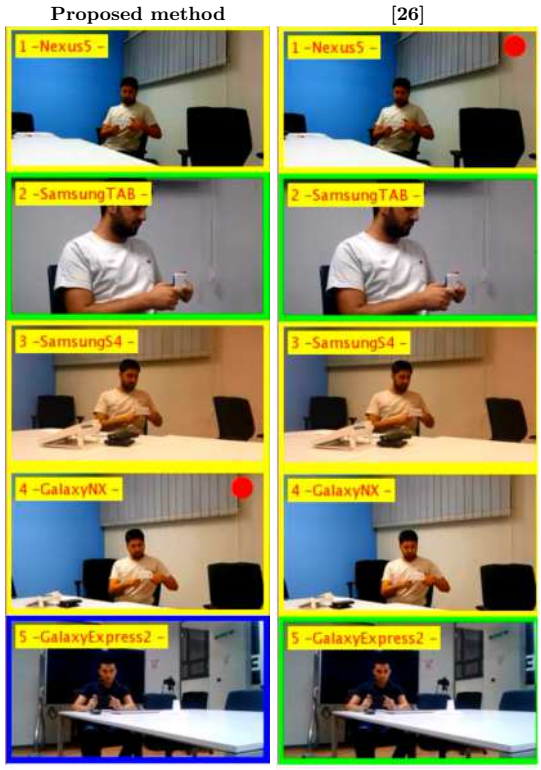

Figure 8: Examples of clustering performed by our system and the approach in [26]. Each column shows the input frames taken from different devices. The color of the border of each frame identifies the cluster. The proposed method produces better results in terms of outlier detection.

6. CONCLUSIONS AND FUTURE WORKS

This work propose a framework to segment and organize a set of egocentric videos for daily living monitoring. We built our system taking inspiration from the RECfusion approach [26]. The experimental results show that the proposed system outperforms [26] in terms of segmentation accuracy and computational costs, both on wearable and mobile video datasets. Furthermore, we tested our approach to solve the popularity estimation problem defined in [26] obtaining an improvement in terms of outliers rejection. Future works can consider an extension of the system to perform recognition of contexts from egocentric images [7] [8] [11] [33] and to recognize the activities performed by the user [9], after the unsupervised organization.

7. REFERENCES

[1] P. Agrawal, R. Girshick, and J. Malik. Analyzing the performance of multilayer neural networks for object recognition. In *Computer Vision–ECCV 2014*, pages 329–344. Springer, 2014.

[2] S. Bell, P. Upchurch, N. Snavely, and K. Bala. Material recognition in the wild with the materials in context database. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[3] V. Buso, L. Hopper, J. Benois-Pineau, P.-M. Plans, and R. Megret. Recognition of activities of daily living in natural "at home" scenario for assessment of alzheimer's disease patients. In *2015 IEEE International Conference on Multimedia & Expo Workshops (ICMEW),*, pages 1–6. IEEE, 2015.

[4] D. Damen, T. Leelasawassuk, O. Haines, A. Calway, and W. W. Mayol-Cuevas. You-do, i-learn:

Discovering task relevant objects and their modes of interaction from multi-user egocentric video. In *BMVC*, 2014.

[5] L. Deng. A tutorial survey of architectures, algorithms, and applications for deep learning. *APSIPA Transactions on Signal and Information Processing*, 3:e2, 2014.

[6] J. Domke and Y. Aloimonos. Deformation and viewpoint invariant color histograms. In *British Machine Vision Conference*, pages 509–518, 2006.

[7] G. M. Farinella and S. Battiato. Scene classification in compressed and constrained domain. *Computer Vision, IET*, 5(5):320–334, 2011.

[8] G. M. Farinella, D. Ravì, V. Tomaselli, M. Guarnera, and S. Battiato. Representing scenes for real-time context classification on mobile devices. *Pattern Recognition*, 48(4):1086–1100, 2015.

[9] A. Fathi, A. Farhadi, and J. M. Rehg. Understanding egocentric activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 407–414. IEEE, 2011.

[10] G. Finlayson, S. Hordley, G. Schaefer, and G. Y. Tian. Illuminant and device invariant colour using histogram equalisation. *Pattern recognition*, 38(2):179–190, 2005.

[11] A. Furnari, G. M. Farinella, and S. Battiato. Recognizing personal contexts from egocentric images. In *Proceedings of the IEEE International Conference on Computer Vision Workshops*, pages 1–9, 2015.

[12] A. Furnari, G. M. Farinella, and S. Battiato. Segmenting egocentric videos to highlight personal locations of interest. In *CVPR 4th Workshop on Egocentric (First-Person) Vision.*, 2016.

[13] Y. Gaëstel, S. Karaman, R. Megret, O.-F. Cherifa, T. Francoise, B.-P. Jenny, and 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 (AAICAD) 2011*, page S245.

[14] C. Gurrin, A. F. Smeaton, and A. R. Doherty. Lifelogging: Personal big data. *Foundations and trends in information retrieval*, 8(1):1–125, 2014.

[15] J. Hosang, M. Omran, R. Benenson, and B. Schiele. Taking a deeper look at pedestrians. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[16] Y. Hoshen, G. Ben-Artzi, and S. Peleg. Wisdom of the crowd in egocentric video curation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 573–579, 2014.

[17] T. Kanade. Quality of life technology [scanning the issue]. *Proceedings of the IEEE*, 100(8):2394–2396, 2012.

[18] N. Kapur, E. L. Glisky, and B. A. Wilson. External memory aids and computers in memory rehabilitation. *The essential handbook of memory disorders for clinicians*, pages 301–321, 2004.

[19] S. Karaman, J. Benois-Pineau, V. Dovgalecs, R. Mégret, J. Pinquier, R. André-Obrecht, Y. Gaëstel, and J.-F. Dartigues. Hierarchical hidden markov model in detecting activities of daily living in wearable videos for studies of dementia. *Multimedia tools and applications*, 69(3):743–771, 2014.

[20] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[21] M. L. Lee and A. K. Dey. Lifelogging memory appliance for people with episodic memory impairment. In *Proceedings of the 10th international conference on Ubiquitous computing*, pages 44–53. ACM, 2008.

[22] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2015.

[23] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[24] Z. Lu and K. Grauman. Story-driven summarization for egocentric video. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2013.

[25] F. L. M. Milotta, S. Battiato, F. Stanco, V. D'Amico, G. Torrisi, and L. Addesso. RECfusion: Automatic scene clustering and tracking in video from multiple sources. In *EI – Mobile Devices and Multimedia: Enabling Technologies, Algorithms, and Applications 2016*. IS&T, 2016.

[26] A. Ortis, G. M. Farinella, V. D'Amico, L. Addesso, G. Torrisi, and S. Battiato. Recfusion: Automatic video curation driven by visual content popularity. In *short proceedings of ACM Multimedia*, 2015.

[27] J. Pinquier, S. Karaman, L. Letoupin, P. Guyot, R. Mégret, J. Benois-Pineau, Y. Gaëstel, and J.-F. Dartigues. Strategies for multiple feature fusion with hierarchical hmm: application to activity recognition from wearable audiovisual sensors. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 3192–3195. IEEE, 2012.

[28] Y. Poleg, C. Arora, and S. Peleg. Temporal segmentation of egocentric videos. In *CVPR*, 2014.

[29] O. Russakovsky, J. Deng, H. Su, J. Krause, S. Satheesh, S. Ma, Z. Huang, A. Karpathy, A. Khosla, M. Bernstein, A. C. Berg, and L. Fei-Fei. ImageNet Large Scale Visual Recognition Challenge. *International Journal of Computer Vision*, 115(3):211–252, 2015.

[30] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *CoRR*, abs/1409.1556, 2014.

[31] B. Soran, A. Farhadi, and L. Shapiro. Generating notifications for missing actions: Don't forget to turn the lights off! In *The IEEE International Conference on Computer Vision (ICCV)*, December 2015.

[32] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[33] B. Zhou, A. Lapedriza, J. Xiao, A. Torralba, and A. Oliva. Learning deep features for scene recognition using places database. In *Advances in Neural Information Processing Systems*, pages 487–495, 2014.