Personalized Egocentric Video Summarization of Cultural Tour on User Preferences Input

Patrizia Varini, Giuseppe Serra, and Rita Cucchiara, *Member, IEEE*

***Abstract*—In this paper, we propose a new method for customized summarization of egocentric videos according to specific user preferences, so that different users can extract different summaries from the same stream. Our approach, tailored on a cultural heritage scenario, relies on creating a short synopsis of the original video focused on key shots, in which concepts relevant to user preferences can be visually detected and the chronological flow of the original video is preserved. Moreover, we release a new dataset, composed of egocentric streams taken in uncontrolled scenarios, capturing tourists cultural visits in six art cities, with geolocalization information. Our experimental results show that the proposed approach is able to leverage user's preferences with an accent on storyline chronological flow and on visual smoothness.**

***Index Terms*—Computer vision, feedforward neural networks, knowledge discovery.**

I. Introduction

In RECENT years the widespreading use of wearable cameras allows capturing everyday life activities such as sport, education, social interactions and cultural heritage visits. However, these videos typically consist of long streams of data with a ceaseless unstable appearance and frequent changes of observer's focus. Moreover, differently from the edited videos, they lack of cuts to separate different video segments. Therefore, there is a urgent demand of adequate tools to process the semantic video content in order to select and store only the meaningful sequences.

Based on this demand, a large number of different egocentric video summarization techniques has been proposed recently. Several approaches have explored low-level, mid-level and high-level features [1]–[3]. Low-level features, typically related to visual motion and color, are able to identify the most stable and homogeneous sequences; the mid and high-level ones [4]–[6] are useful to identify salient content such as social interactions or objects manipulations. Furthermore, multimedia approaches have also investigated physiological data such as brain waves [7] and gaze [8] to select the relevant segments.

Although these summarization techniques deal with egocentric characteristics they produce a unique summary. The generation of a single summary could be adequate in some controlled domains such as video surveillance of a specific area [9], [10], in which the salient events are generally predefined. This assumption can not be exploited in egocentric videos captured in unconstrained scenarios. In this case users can have different preferences. Therefore, they might prefer retaining in the summary some events rather than others. For instance, art lovers and fashion enthusiasts might desire to extract different summaries from their egocentric videos taken during a visit in an art city. In addition, a user may want to share personalized summaries with different people, such us parents or friends.

In this paper we address this challenge. In particular, we propose a summarization technique in the cultural heritage domain that is able to generate customized summaries according to specific user preferences. To identify candidate relevant sequence in an egocentric video, we first present a behavior pattern classifier based on a novel Convolutional Neural Network (CNN) architecture exploiting apparent, 3D motion and visual assessment features. Based on these detected sequences, the proposed summarization approach selects the items considering three key elements: the attention behavior of the wearer, the visual and semantic coherence with the user preferences and a narrativity grade gauged by a new proposed metric.

To assess sequences related to the user preference, we propose to build a set of visual classifiers. Since preferences are potentially limitless, semantic classifiers are built on-fly using a data-driven approach, that exploits geolocalization and DBPedia semantic knowledge. Moreover, to recover the chronological flow of the most relevant sequences we propose a Personalized Page Rank approach that takes into account narrativity metric assessed by means of importance and relevance into the frame sequence.

Finally, we present several experiments on a new dataset of unconstrained egocentric videos and we demonstrate the effectiveness of our solution. Fig. 1 shows an example of three different summaries: an uniform summary no preference driven, and two customized summaries obtained with the proposed method, using two sets of different user preferences.

To sum up, our contribution can be summarized as follows:

1) we propose a novel user behavior classifier, based on a 3D CNN approach, that joints apparent, 3D real motion features and visual quality features. The proposed ap-

Manuscript received December 2, 2016; revised March 28, 2017; accepted May 2, 2017. Date of publication May 18, 2017; date of current version November 15, 2017. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Tao Mei. (*Corresponding author: Giuseppe Serra*.)

P. Varini and R. Cucchiara are with the Department of Engineering "Enzo Ferrari," University of Modena and Reggio Emilia, Modena 41100, Italy (e-mail: patrizia.varini@unimore.it; rita.cucchiara@unimore.it).

G. Serra is with the Department of Mathematics, Computer Science, and Physics, University of Udine, Udine 33100, Italy (e-mail: giuseppe.serra@uniud.it).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2017.2705915

Fig. 1. Different summaries produced by the same original video, one uniform summary and two customized summaries based on our method according to two different sets of user preferences, "Sport Cars, Romanesque" (related frames edges surrounded respectively in blue and green) and "Shopping, Fashion" (related frames edges surrounded in red).

proach, that achieves state of the art results, deals with the egocentric challenge of detecting attentive behavior pattern also in presence of video shaking appearance;

2) to identify the sequences related to user preferences we present a new data-driven approach based on a set of classifiers dynamically learned. Differently from previous approaches, the proposed approach exploits jointly DBpedia Knowledge Base and geolocalization of the user to extract reliable positive and negative samples;
3) we define a new technique to recover in the summary the original chronological storyline progression of the original video. This deals with the egocentric issue to recover a meaningful narrative thread, from the bothered jumping and discontinuous sequence of scenes with ever changing observer's focus.
4) we publicly release a new unconstrained dataset containing egocentric videos of cultural heritage visits in six art cities. To best of our knowledge, this is the first egocentric video dataset that contains jointly geolocalization information, behavior classification, narrative milestones.

II. Related Work

Video summarization has been widely studied in different application domains such as sport [11]–[13], surveillance [14], [15], multimedia retrieval [16], [17]. Effective approaches are based on high-level features, such as salient people, objects and actions [18] and more recently on deep learning techniques such as Long-Short Term Memory [19], [20]. Differently from the third-person approaches, egocentric vision summarization, only recently addressed by research community, requires new techniques to deal with continuous changes of observers focus and with lack of hard cuts between scenes [21], [22].

A. Low-, Medium-, High-Level Features-Based Approaches

To address egocentric challenges, several approaches have explored the usage of low-level and high-level features such as color, visual motion and presence of people or objects. Lu *et al.* [1] presented a technique based on optical flow and bluriness features to divide videos into sub-shots. Subsequently, they selected a chain of sub-shots choosing the ones where reciprocal influence between important objects and characters was detected.

Gygli *et al.* [5] identified video segments by using a "superframe" segmentation on different raw videos, then they scored visual interestingness per superframe using a set of low-features (color, edges, blur), mid-features (landmarks) and high-level concepts (recognition of faces, objects using segmentation). Based on these scores, they selected an optimal subset of superframes to create a maximum informative synopsis. Moreover, Gygli *et al*. [23] formulated the task of video summarization as a subset selection problem. In particular, they learned a linear combination of submodular functions, using structured learning with a large-margin formulation, directly from reference summaries created by human annotators.

Lidon *et al.* [24] first selected informative images by a two-class CNN. Then, relevance ranking was obtained by integrating techniques for saliency detection, object recognition and face detection. Poleg *et al.* [3] learned to classify camera's wearer motion leveraging a three dimensional CNN architecture, fed with stream sparse optical flow volumes as input. In addition, Poleg *et al.* [2] proposed a fast-forwarding approach based on motion related features that obtained timelapses or hyperlapses by selecting optimal set of frames avoiding non-aligned consecutive sampled frames.

Joshi *et al.* [25] obtain hyperlapses by choosing frames from the original video that best match a certain speed-up, surreptitiously achieving camera motion smoothing. Okamoto *et al.* [26] starting from first person videos produces short walking route guides in which they dynamically control video playing speed. This is achieved assessing egocentric motion features assigning different importance score to the sequence of the streams in which appear urban items such as pedestrian crosswalks or traffic lights. Lin *et al.* [27] propose a Structured SVM detector to detect highlights analyzing the context of each video sequence. Yao *et al.* [28] learned temporal and spatial highlights from appearance of video frames and temporal dynamics across frames using late fusion of two-streams spatio-temporal CNN structure. Zhang *et al.* [20] investigated how

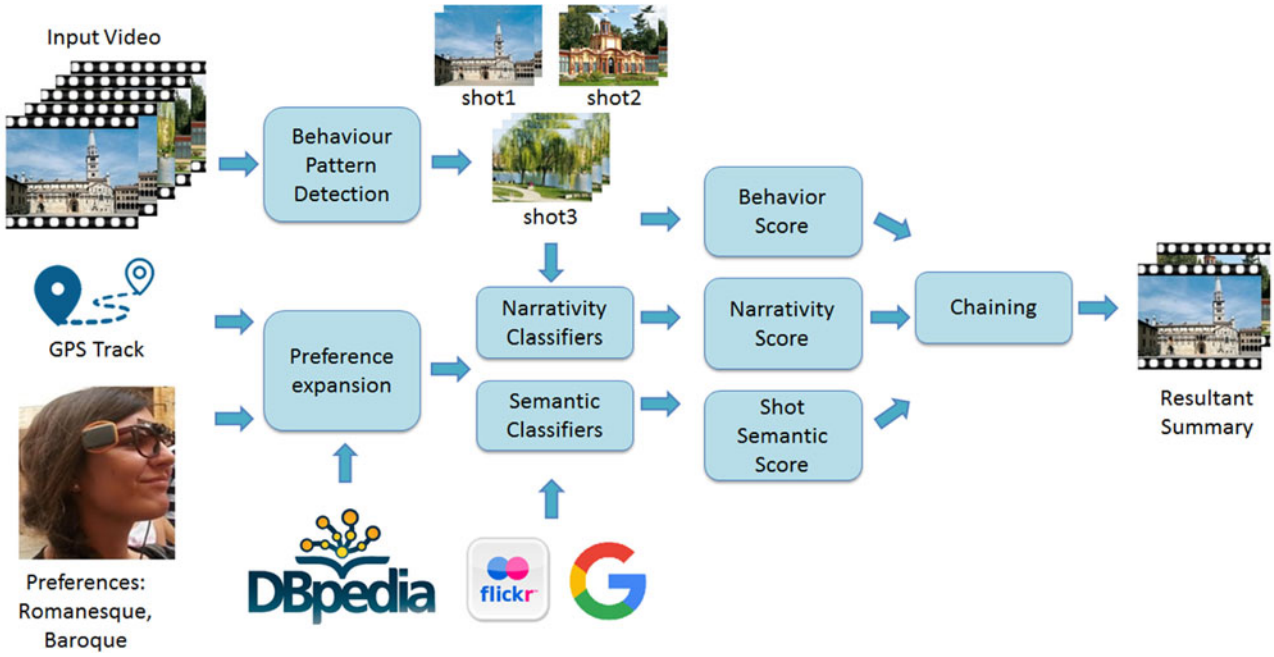

Fig. 2. Schematization of the proposed method: maximization of objective function is based on visitor behavior, and semantic and narrativity scores are computed on detected shots to obtain the final summary.

to apply Long Short-Term Memory model to supervised video summarization.

B. Sensor-Based Approaches

To improve the egocentric summaries accuracy, other approaches presented solutions based on information extracted from heterogeneous sensors. For instance, Ng *et al.* [29] proposed an egocentric video summarization method that used physiological data such as brain waves to select the relevant segments of video. Lee *et al.* [4] proposed an approach that focused on the detection of important cues for each frame, such as objects and people the camera wearer interacts with. To identify these frames they proposed to combine wearers gaze and visual features, such as objects' motion and appearance. More recently, Xu *et al.* [8] used gaze information to identify the skims and to score their relevance, subsequently maximized reciprocal shots information in order to achieve a summary in which relevance, diversity, fidelity and compactness were optimized.

C. Personalized Summarization

Although these above mentioned egocentric summarization approaches are valuable, they miss a key element of the egocentric paradigm: the person's specific intention. In fact, everyone is different thus there does not exist a univocal summary which is the best one for everybody.

A recent approach by Sharghi *et al.* [30] proposes a generalscope video summarization approach where shots are selected according to relevance to input queries and importance in the stream context, leveraging a Sequential and Hierarchical Determinantal Point Process.

An early version of the approach proposed in this paper was introduced in a preliminary work [31] specifically designed for the egocentric domain, in which the behavior of the camera's wearer is directly recognized. Here, we propose new key components in order to achieve more reliable results. Differently from our previous work, we present a new algorithm for Behavior Pattern Recognition that is able to achieve robust results even in unconstrained scenarios. We propose a more formal approach, based on DBpedia, for the semantic expansion and filtering of the preferences specified by the user. We introduce a new solution to preserve the chronological flow in the summary. Finally, we create, and contextually publicly release, a larger egocentric video dataset of Cultural Heritage tour visits.

III. Personalized Egocentric Video Summarization

Our solution takes a long first-person video of cultural or touristic experience and a text expressing user preferences as input and returns a customized short video summary as output. Our goal is to preserve logical and chronological storyline in which we want to put in evidence the segments related to the preferences expressed by a particular user. Fig. 2 synthesizes our approach.

In egocentric videos of cultural heritage experiences, the presence of relevant information is likely to be put in relation with camera's motion behavior patterns more than others. For example, an interesting behavior "looking at an artistic monument" is characterized by some little movements of the head for an interval of time sufficient to establish a visual fixation pattern [32], probably standing still for a while. This behavior has a very different pattern of motion with respect to other less interesting behaviors, like wandering around or walking straight that show wide movements of the head.

Therefore, we define a set of motion behavior classifiers based on real, apparent motion and quality assessment of the visual

content to extract homogeneous skims of the video (from now on we call them "ego-shots" - see Section III-A for more details). Using behavior detection we are also able to assign a score to ego-shots according to the observer's behavior pattern: we assign a high score to ego-shots with the presence of user attention and lower scores to ones containing less relevant activities such as wandering, walking, running, etc.

However, recognizing the presence of attention behavior pattern alone is not sufficient. In fact, actions like "looking at the phone", "waiting to cross a road", "looking at bus timetable", may all be put in relation with attention motion behavior pattern, but they are generally uninteresting in this scenario. Consequently, we introduce in our solution two other key aspects: "semantic relatedness" with the user preferences and "narrative importance" to recover original storyline.

Semantic relatedness is achieved building a set of visual classifiers based on the user preferences. However, since topics of interest are potentially unlimited, classification based on a number of predefined rigidly defined classes may lead to poor performances. Thus, we build specific classifiers for the topics of interest using a data-driven approach that extracts on-fly reliable training images exploiting DBPedia semantic knowledge and GPS tour information.

Narrative importance has the target of recovering the chronological flow of most salient events. For this purpose we want to identify the pivotal skims for each sequence in the narration, as the segments with the highest amount of information.

Let V be the original video regarded as the composition of a set of consecutive shots $\{S_1, S_2, \ldots, S_n\}$, the personalized video summarization is arranged in chronological order, selecting the ego-shots S^* that maximize the objective function depending from Behavior score $B(S)$, User Preference Semantic Score $P(S)$ and Narrative Importance score $N(S)$

$$O(S^*) = \underset{S_i \in V}{\text{Arg Max}} \quad (B(S_i) + P(S_i) + N(S_i)). \qquad (1)$$

It should be noted that we could adopt a weighted sum between the three scores in order to gauge the contributions. However, while a optimization of these weights could slightly improve the performance, it would lead to a fine tuning too excessively specific a in contrast with our purposes.

A. Behavior Pattern Detection

Tailored on a typical cultural experience scenario, we assume the hypothesis that shaped behavior of visitors motion is a relevant information to identify and classify interesting or irrelevance sequence of the video. Thus, we first segment a video into ego-shots characterized by homogeneous behavior patterns. In detail, we define six behavior pattern classes: "Attention" (where user is paying attention to something), "Looking around" (searching for something etc), "Walking", "Running", "On wheels", "Wandering".

To detect these activities we propose a CNN with 10 layers trained on motion and frame visual assessment features (see Fig. 3). Motion features encode both visual apparent and real components. Visual motion feature is based on optical flow estimated using the Farneback algorithm [33].

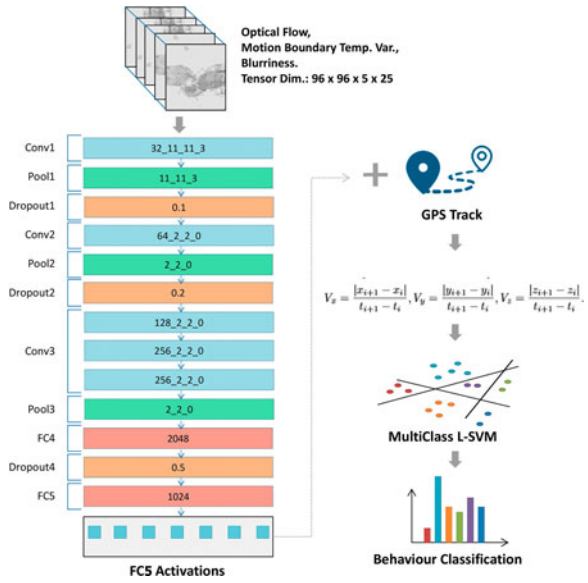

Fig. 3. Behavior classification schema. Dropout layers are used only during the training phase. Convolutional layer parameters are expressed as $d\_f\_s\_p$, where d is the number of filters, $f \times f$ is the kernel size, p is their depth, and s the stride applied to the input. Pooling layers parameters are expressed as f_s_p, where $f \times f$ is the kernel size, s is the stride, and p the depth.

In particular, considering the Optical Flow computed for each couple of consecutive frames, we extract the relative apparent velocity and acceleration gradient for both horizontal and vertical components for each pixel, obtaining four bi-dimensional tensors shaped as the frame size.

Real motion features are obtained collecting the geographic locations of the visitor by means of a GPS sensor. Locations are represented as a temporal sequence of the spatio-temporal points, meant as pairs compound with coordinate in space and in time $\{p_0 = (x_0, y_0, z_0, t_0), ..., p_N = (x_N, y_N, z_n, t_N)\}$, where $(x_i, y_i, z_i) \in \mathcal{R}^3, t_i \in \mathcal{R}^+$ for $i = 0, 1, ..., N$ and $t_0 < t_i < t_N$. Given them, we extract three components of real 3D velocity of visitor as

$$V_x = \frac{|x_{i+1} - x_i|}{t_{i+1} - t_i}, V_y = \frac{|y_{i+1} - y_i|}{t_{i+1} - t_i}, V_z = \frac{|z_{i+1} - z_i|}{t_{i+1} - t_i}. \qquad (2)$$

To assess the frame quality, we compute frames blur feature by using the method in [34]. This approach assumes that the sharpness of a frame is contained in its gray component and estimate the blur annoyance only on the luminance component, computing and evaluating the line and row difference between the original image and the image obtained applying to it a horizontal and a vertical strong low-pass filter.

The blur feature employed in our model is the absolute difference between the original image and the filtered image, averaged on each couple of frames, and is shaped as the frame size.

Apparent motion and blur descriptors are resized to 96×96 pixel and stacked to obtain our overall apparent motion descriptor for a couple of frames: a tensor shaped $96 \times 96 \times 5$. To take into account time dimension the tensor descriptors are further stacked over one second that provides a final descriptor of shape $96 \times 96 \times 5 \times 25$ (resampling to 25 fps videos whose frame rate is different).

Our network can be described in shorthand notation as C1(64; 11; 3; 3)-P1-C2(128; 5; 2; 1)-P2-C3-1(256; 3; 2; 1)-C3-2(384; 3; 2; 1)-C3-3(256; 3; 2; 1)-P5-FC4(2048)-FC5(1024), with 3 dropout layer, 0.1, 0.3, 0.5, during training phase. In the above notation C(d, f, s, p) indicates a convolutional layer with d filters, kernel size $f \times f$, depth p, applied to the input with stride s. The first convolutional layer and pooling layer are implemented as 3D, while following ones are standard 2D. FC(z) is a fully connected layer with z nodes. Rectified Linear Units (ReLU) is employed. We adopt the L2-SVM loss to train the CNN [35].

Once the aforementioned CNN is trained from scratch on the training set, FC5 activations are extracted and concatenated with real motion features, after L2 normalization. A linear SVM is then used to classify shots of the egocentric sequences, on 1027 dimensional feature vectors. Fig. 3 presents a schematization of our behavior classifiers.

B. Semantic Classifiers Based on User Preferences

To score the semantic relatedness of the ego-shots to the User Preferences, we build a recognition system based on visual classifiers. Since topics of interest requested by the user can be potentially limitless, we propose a data-driven approach that gathers positive and negative training images from the web.

In our solution, for each video the user can indicate multiple preferences in the form of a set of sequences of words. Given a user preference a pre-processing step to eliminate stop words, dates, numbers, punctuations and symbols, is performed; it is converted in the lemmatized form. Thus, a single user preference can be represented as a set of lemmatized concepts $C = \{c_1, c_2, ..., c_n\}$.

To extract positive and negative training images from the web, it is necessary to define a set of adequate text input queries. In fact, a mere query using only the concatenation of user preference terms might achieve poor results due to language and semantic ambiguity.

Since Cultural Heritage items are often highly location-specific and since some concepts and semantic relationships may be unknown even to the user, we show that semantic exploration on a reliable Knowledge Base (KB) and user localization (L) can improve the search terms to collect training images.

To achieve this target, we choose to exploit DBpedia Knowledge Base, because it is an important and constantly updated dataset in Linked Open Data, that consists of semantic structured content (hyponyms, hypernyms, synonyms, antonyms) with cross references between related topics [36].

Dbpedia can be represented as an undirected graph $KG = \{V, E\}$ where V are the nodes representing stemmed and lemmatized concepts and $E \subset V \times V$ are the edges representing the links among nodes.

For each of concept c_i of a user preference we attempt to match it with Dbpedia graph KG using a text match. In the case of exact match, there might be the need of disambiguation with nodes of KG. If disambiguation is needed, we choose the node that minimizes the sum of Dijkstra distances [37] between the ambiguous node and nodes related to other concepts of C and location L.

In this way, it is possible to associate concepts in C to KG nodes $K = \{k_1, k_2, ..., k_m\}$, which we call "Seed nodes" (notice $m = n$ in the case in which all the concepts are matched to KG).

Therefore, a subgraph from KG considering only the set of nodes K and their edges is built. This subgraph is enriched recursively taking into account, for each k_i, all the nodes connected by edges related to Cultural Heritage domain (i.e., type skos:broader, skos:broaderOf, rdfs:sub-ClassOf, rdfs:type and dcterms:subject).

For what concerns the number of recursive steps, based on our experiment we found that a distance of two-hops is optimal in most cases. In fact, node expansions to three hops usually produce very large graphs and introduce a considerable amount of noise (see Section IV for more details). Concepts in C, that are not linked to any nodes of KG, are not expanded. For example, let's consider the simple user preference $C = \{$"Romanesque", "Church" $\}$; the corresponding set of seed nodes on the KG is $K = \{$"Church","Romanesque Architecture"$\}$.

Once obtained the graph, to minimize the noise induced by concepts less related on the user preference, we detect and analyze the semantic communities to identify the most connected one centered on K nodes.

For this purpose we adopt the recursive Girvan-Newman algorithm [38]. This starts with computing the "betweenness" score for each of the edges ("betweenness" of an edge is the number of shortest paths between pairs of nodes that run along it). Then edges with the highest score are removed and the betweenness of all edges affected by the removal are computed. The last two steps are repeated until no edges remain.

We use graph metrics, to assess how focused and how dense are the detected communities around the expanded user preferences. In particular, Focused Current Flow Betweenness Centrality [39] (FCFBC), Average Node Connectivity [40] (AVC), Focused Information Centrality [41] (FIC) and Average of the shortest paths between community nodes and location L (computed using Dijkstra algorithm), are computed for the member nodes (and relative edges) of each detected community, in order to evaluate density of the communities and their semantic distance from the location.

At last starting from $K = \{k_1, k_2, \ldots, k_m\}$ we identify a new set of concepts $\hat{K} = \{k_1, k_2, \ldots, k_m, k'_1, k'_2, \ldots, k'_p\}$, where $k'_1 ... k'_p$ are the additional nodes of the detected community.

The Fig. 4 shows the graph obtained by recursively expanding seed nodes of K and the detected communities. For sake of clarity we show only a bunch of nodes since the cloud of connected nodes is much larger.

Starting from $\hat{K}$, we download images and their metadata from Visual Online Repositories like Flickr, Google Photos, leveraging textual queries to extract 2000 positive and negative samples for each user preference.

The query to extract a set of positive image samples P, is obtained by combining in "OR" each element. For each image in P we extract a set of metadata, which consists of tags and/or description (preprocessed eliminating stop words and applying stemming).

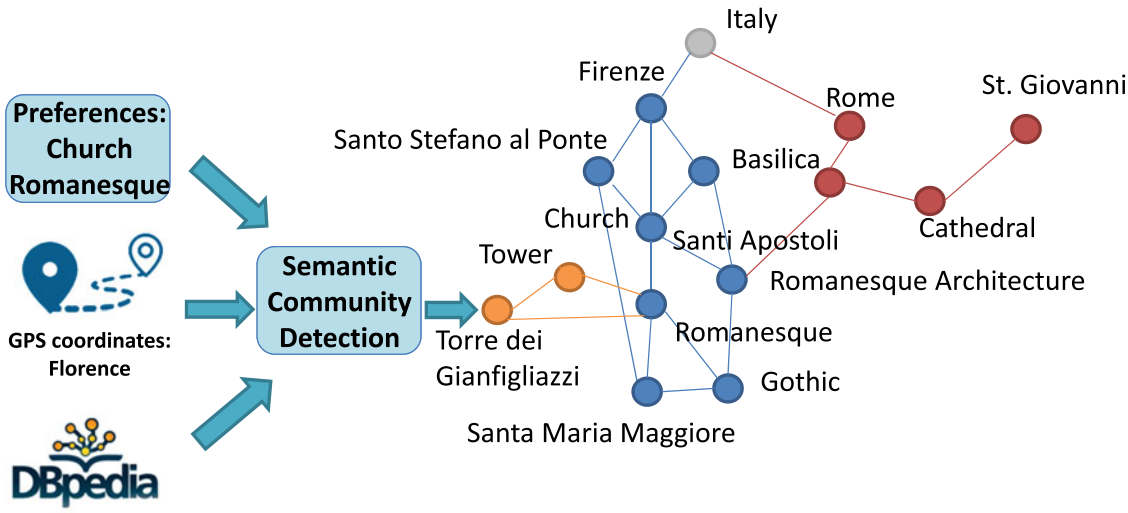

Fig. 4. Schematization of the semantic expansion approach: the blue graph is the community with the maximum connectivity and centrality of the expanded graph; the yellow and red communities contain the nodes linked with the Seed nodes K but are less connected and less focused than the blue one.

The query, to extract negative samples, is generated by combining in "NOR" the set of hypernyms of $\hat{K}$ nodes.

To further reduce noise and increase reliability of our training sample images we leverage jointly visual and semantic features extracted by textual metadata (tags and descriptions). To extract and encode visual content of each image, we use the convolutional network VGGNET-16, pre-trained on Imagenet, that has shown reliable results in image classification [42]. In particular, we use the VGGNET-16 as a feature extractor, removing the output layer and using the activations of the last fully connected layer, after applying L_2 normalization, obtaining at last 4096 dimensional feature vectors.

Semantic representation features are extracted from downloaded image tags and descriptions using the Word2Vec algorithm [43]. We train Word2Vec model on "Turtle" dump file of DBPedia english short abstracts, containing 4.3 M triples, setting the dimensionality of our feature vector to 100. For each preprocessed metadata of a downloaded image I we compute its representation in the semantic vector space; then the final semantic feature representation $\mathbf{w}_I$ of I is obtained by element-wise addition of their vector representations.

To filter the most noisy candidate training images, we discard the candidates that are semantically too far from the expanded concepts $\hat{K}$. To compute their semantic distance, we map $\hat{K}$ in the semantic vector representation $\mathbf{w}_{\hat{K}}$ following the same approach as described above. Thus, semantic distance of I from original $\hat{K}$ concepts, is obtained by computing the cosine distance between $\mathbf{w}_I$ and $\mathbf{w}_{\hat{K}}$. Images with a semantic distance below a threshold (empirically, ones with less than 10% of the average) are discarded.

For the non discarded images, we concatenate their metadata semantic representation to visual feature vector, obtaining a joint visual and semantic descriptor, whose dimension is 4196. Then, we perform unsupervised agglomerative clustering (DBSCAN, with dynamically adaptively computed parameters [44]) on the final features, discarding the images belonging to clusters with a number of elements less than 3% of the original image set. The non discarded images can now be used to build for each user preference a visual classifier. At last, for each shot we compute the User Preference Semantic Score $P(S)$ obtained averaging visual classification score over the shot length.

C. Shot Narrativity Score

At this point, according to (1), we want to assign a score that quantifies narrative relevance of each ego-shot, i.e., the capacity of a chosen subset of video segments to represent an important milestone in the chronological flow of events, with a focus on relevance and diversity. In fact, let's for instance assume that a visitor is walking for a while, far from any Points of Interest; POI detectors give to these scenes very low scores, but their complete elimination might generate a hole in the chronological flow. Recent approaches pursue mutual information maximization between original video and summary [8] assessing shots information diversity with methods like maximum margin relevance [23] or determinantal point process [20], eventually using people or objects as plot pivots [24]. Unlike egocentric videos of daily activities (a food preparation, object manipulation etc.) or videos taken by police body-worn cameras, typical Cultural Heritage tour egocentric videos consist of long sequences of events in hugely unconstrained scenarios. Thus, maximizing mutual information between summary and video or maximizing visual diversity of skims, lead to poor results, because a number of irrelevant but visually rare skims may obtain a passing score and might be included in the summary. Differently, we are interested in identifying the story most meaningful and influential joints in terms of "Relevance" and "Diversity".

To address this issue, we start regarding the original video V as a directed graph, where ego-shots $\{S_1, S_2, \ldots, S_n\}$ are represented by nodes, edges and their weights are dependent on Relevance within a frame of time (see for more details Section III-C.1) and we then compute nodes centrality to assess their representativity. In fact, nodes centrality is widely

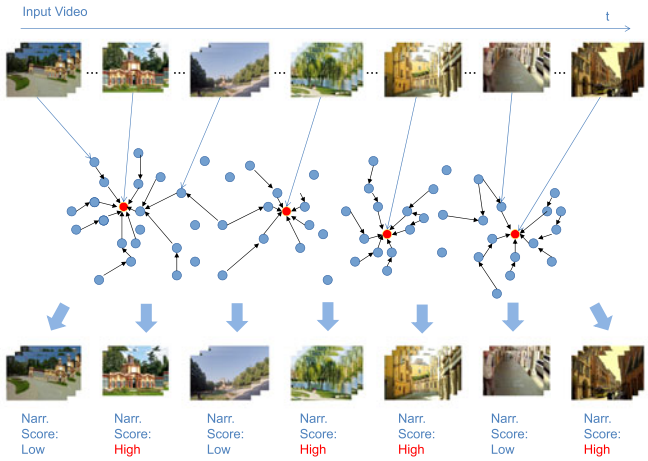

Fig. 5. Narrativity detection with personalized page rank approach based on visual similarity and visual relevance. Narrativity score here is denoted as low or high, instead of a numeric value, to make the example more readable.

used to select the most important nodes in terms of influence in a graph.

In particular, we propose to use the Personalized Page Rank (PPR) approach [45] to compute centrality. The Personalized Page Rank formulation yields a node centrality measure on weighted directed graphs, that depends on the number of received inbound links, the centrality of the linkers and, inversely, on their link tendency (the more lavish, the less evaluated their contribution). In matrix form, the formulation for node centrality score can be written as

$$\mathtt{C} = [d\mathtt{U} + (1-d)\hat{\mathtt{W}}]^{\mathrm{T}}\mathtt{C} \tag{3}$$

where matrix $\hat{\mathtt{W}}$ is obtained from the weighted graph adjacency matrix $\mathtt{W}$ dividing by row sum: $\hat{W}(u,v) = \frac{W(u,v)}{\sum_{z \in adj(v)} W(z,v)}$, $\mathtt{U}$ is a square matrix with all elements being equal to $1/n$, being n the total number of nodes in the graph (i.e., the number of ego-shots), and d is a "damping factor" fixed to 0.15 (based on preliminary results). Let's be $\mathtt{Z} = [d\mathtt{U} + (1-d)\hat{\mathtt{W}}]$ (3) can be written as

$$\mathtt{C}^{\mathrm{T}} = \mathtt{Z}\mathtt{C}^{\mathrm{T}}. \tag{4}$$

This formulation shows that $\mathtt{C}^{\mathrm{T}}$ is the left eigenvector of the matrix $\mathtt{Z}$ with the corresponding eigenvalue of 1. Based on Perron-Frobenius theorem [46] this eigenvector exists and can be uniquely computed. We solve this eigenvector equation using the Power Iteration method [47].

Now, sorting in descending PPR order the graph nodes, we re-score them penalizing their temporal proximity. At the end of the ranking process, the score values are normalized. The Fig. 5 synthesizes this approach.

1) Weighted Adjacency Matrix Definition: we define edge weights between S_i and S_j as

$$w_{i,j} = e^{-\frac{\Delta(N_{i,j})^2}{\tau}} \cdot P(S_i, S_j). \tag{5}$$

The term $e^{-\frac{(\Delta N_{i,j})^2}{\tau}}$ takes into account temporal distance between shots. In particular, $\Delta N_{i,j}$ is the difference in seconds between the shots S_i and S_j, and τ is a normalization factor (fixed empirically based on preliminary experiments). The factor $P(S_i, S_j)$ encodes visual similarity and relevance between two shots. We define it as

$$P(S_i, S_J) = VS(S_i, S_j) \cdot \max(VR(S_i), VR(S_j)) \tag{6}$$

where $VS(S_i, S_j)$ is the function that measures the visual similarity between shot S_i and S_j and $VR(S_i)$ measures the visual relevance of the shot S_i. Edge direction is given by $\frac{R(S_i)-R(S_j)}{|R(S_i)-R(S_j)|}$ except in the case in which $R(S_i) = R(S_j)$; in this case the nodes are linked by two symmetric edges.

We compute visual similarity between ego-shots extracting features from deep 3-dimensional convolutional network named $C3D$ [48] and computing cosine similarity between the feature vectors represented by last dense layer activations, whose dimension is 4096.

To compute visual relevance (VR) we assess three specific visual attributes: Colorfulness, Informativeness, Aesthetic Appearance. Indeed in Cultural Heritage domain, visitors are likely to spend long times engaged in potentially irrelevant activities involving colorless, dull backgrounds (staring at phones, informative boards, timetables), repetitive patterns (asphalt roads, brick walls, pavements, facades), crooked shots (odd shots captured during actions involving wide head movements).

To assess jointly the three attributes we exploit a multi-column CNN [49], that combines various CNN columns.

The single column CNN consists of the aforementioned C3D. Features are extracted from the last dense layer. To exploit the multiple attributes approach, multi-task learning is used [50], where feature representation and classification error minimization are constructed in a joint manner for the three attributes. Assuming we have Informativeness labels $\mathbf{y_I}$, Colorfulness labels $\mathbf{y_C}$ and Aesthetic Appearance $\mathbf{y_A}$ for all training images, the problem could be formulated as minimization of the cross-entropy loss

$$\begin{aligned} L(\mathbf{x}, \mathbf{y}, \mathbf{w}) \\ = -\Bigg(\sum_{i=1}^{N}\sum_{l \in L_C} \mathbf{I}(y_{Ci} = l) \ln p(y_{Ci} = l | x_i, \mathbf{w_C}) \\ + \sum_{i=1}^{N}\sum_{l \in L_I} \mathbf{I}(y_{Ii} = l) \ln p(y_{Ii} = l | x_i, \mathbf{w_I})\Bigg) \\ + \sum_{i=1}^{N}\sum_{l \in L_A} \mathbf{I}(y_{Ai} = l) \ln p(y_{Ai} = l | x_i, \mathbf{w_A})) \end{aligned} \tag{7}$$

where $\mathbf{x}$ is the feature vector of N training images, L_C, L_I and L_A are the label sets. The sets $\mathbf{w}_I = \{\mathbf{w}_{I,l}\}_{l \in L_I}$, $\mathbf{w}_C = \{\mathbf{w}_{C,l}\}_{l \in L_C}$ and $\mathbf{w}_A = \{\mathbf{w}_{A,l}\}_{l \in L_A}$ are the model weights respectively for Colorfulness, Informativeness and Aesthetic Appearance, initialized to the values of the pre-trained single networks. $\mathbb{I}(z) = 1$ if z is true and 0 if false and, for each attribute a in {Colorfulness, Informativeness, Aesthetic Appear-

ance}, probability p_a is expressed as

$$p_a(y_{ai} = l|\mathbf{x_i}, \mathbf{w_{al}}) = \frac{e^{\mathbf{w_{al}^T x_i}}}{\sum_{h \in L_a} e^{\mathbf{w_{ah}^T x_i}}}. \quad (8)$$

At last, for each shot, Relevance $VR(S_i)$ is computed averaging the prediction on the shot frames.

IV. Experimental Results

To evaluate the performance of our approach we test each component of the proposed approach on a novel dataset. We assess separately effectiveness of Behavior Pattern classification, Semantic Relatedness detection and Narrative Importance measurements. Finally, we use a blind taste test to score the summaries obtained using the proposed solution with respect to four competitors, including a recent state-of-the-art approach [23].

A. Art City Egocentric Dataset

We collected 48 videos captured by tourists during visits in six italian art cities: Bologna, Ferrara, Modena, Parma Ravenna, Reggio-Emilia.

They show visitors experiences such as a visit to a cultural heritage site or an attraction (churches, monuments etc), or activities like shopping or walking.

The cameras are placed on the tourist's head. Videos, captured by five different users, are of variable length, ranging from a minimum of 15 minutes to a maximum of 25 minutes. Moreover, they are taken in an uncontrolled setting with a resolution variable from 720×576 to 1920×1080 and a variety of brightness and chromaticity conditions during different daytimes and seasons.

For each video, annotations are added on three different semantic dimensions: observer's behavior, presence of Points or Items of interest and narrativity related informations. A total of 26100 frames is annotated for each semantic dimension. GPS annotations are fully provided for all the videos length.

Granularity of GPS sensor is one second in time and 1 meters linear displacement in space. To the best of our knowledge this is a novel dataset specifically designed to analyze this egocentric scenario.[1]

B. Experiments on Behavior Pattern Classification

First, we examine the effectiveness of our solution for behavior recognition based on frame quality assessment, visual and real motion pattern. A subset of 26100 annotated frames of our dataset is fully annotated with the six behavior pattern classes and used to test our methodology. Each frame is resized to 96×96 pixels, then Dense Optical Flow and Motion Boundary temporal derivative are extracted for each couple of frames. Blur on horizontal and vertical axes are as well combined (i.e., averaged) for each couple of frames. The five tensors are stacked and represent the network input. We use a 10-fold cross-validation approach partitioning the dataset for training, validation and testing, (dimensioned 80, 10 and 10 percentage of the dataset respectively). We compare our technique with recent and related approaches [1], [3], [31]. To be comparable for all of these methods we use the same experimental settings. Table I presents comparative results between our solution and these approaches in terms of average class F1-score. As can be seen, our solution achieves a better performance. This is mainly due to the addiction, differently from previous approaches, of motion boundary and 3D real motion features in our 3D CNN architecture.

TABLE I
Comparison of Classification Accuracy

Method	Accuracy
Lu *et al.* [1]	63.67
Varini *et al.* [31]	74.17
Poleg *et al.* [3]	90.67
Our method	**92.17**

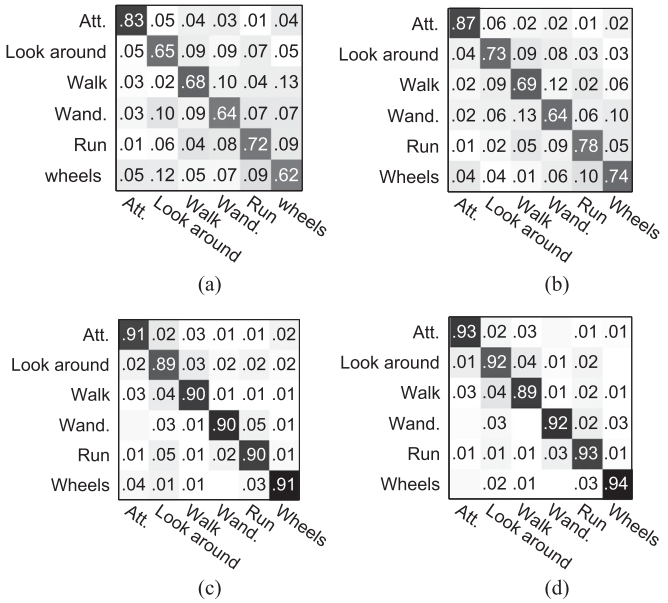

Fig. 6. Classification accuracy using different techniques: (a) Lu *et al.* [1], (b) Varini *et al.* [31], (c) Poleg *et al.* [3], and (d) our approach.

In particular, Fig. 6 shows the confusion matrices for the aforementioned approaches. Adding descriptors related to 3D real motion helps to distinguish higher speed motions; this can be seen in classes such as "Running" and "Moving on Wheels". Furthermore motion boundary features better encode the wide head motion or sprawl, improving the performance of classes such as "Looking around" and "attention" patterns.

C. Experiments on User Preferences Expansion

To test the performance of our semantic expansion of the user preferences we simulate and analyze different user inputs. Table II shows the initial user preferences, the city locations and the expanded preferences. For each expanded item we evaluate the precision (i.e., the number of correct terms with respect to the number of expanded terms). The table also reports the

[1][Online]. Available: http://imagelab.ing.unimore.it/egosummarization

TABLE II
RESULTS OF SEMANTIC EXPANSION OF DIFFERENT USER PREFERENCES. EXPANDED PREFERENCES OBTAINED USING THREE HOPS ONLY SHOW THE ADDITION TERMS WITH RESPECT TO TWO HOPS RESULTS

User preference	Expanded Preferences	Location	Hops	Precision	Avg. Node Connectivity	Avg. Focused Information Centrality
Church, Romanesque	Romanesque, Architecture, Cathedral, Church, Cathedral Torre Civica and Piazza Grande, Gothic, Lanfranco, Painting, Reliefs, Saint Geminianus, Sculpture, Wiligelmus	Modena	2	92	2.25	0.635
Church, Romanesque	Romanesque, Medieval, Catholic, Basilica, Castle, Torre della Ghirlandina, Thomas of Britain, Tristan	Modena	3		1.98	0.462
Architecture, Renaissance	Art, Renaissance, San Pietro, Church, Architecture, Painting,	Modena	2	91	1.81	0.514
Architecture, Renaissance	Art, Renaissance, Catholic, Galleria Estense, Music Tintoretto, Veronese, Correggio, Dell'Abate	Modena	3	72	1.83	0.519
Architecture, Baroque	Baroque, San Agostino, San Barnaba, San Carlo, Chiesa del Voto	Modena	2	95	1.89	0.457
Architecture, Baroque	Baroque, Catholic, Basilica, Reni, Lana, Guercino, Velasquez, Bernini	Modena	3	97	1.88	0.523
Sport, Cars	Sportcars, Enzo Ferrari, Ferrari, Maserati, Lamborghini, de Tomasi, Pagani, Automobile, Manifacturer, Museum	Modena	2	99	1.75	0.467
Sport, Cars	Sportcars, Racecar Driver, Race, Rally, Team	Modena	3	90	1.74	0.433
Fashion, clothing, Shopping	Fashion, Clothing, Shopping, Stores, Online Shopping, Window Shopping, Mall	Modena	2	86	1.83	0.423
Fashion, Clothing, Shopping	Fashion, Clothing, Shopping, Industry, Activity, Retail	Modena	3	62	1.83	0.396
Renaissance, Architecture	Baroque, Architecture, Bentivoglio, Certosa, Church of Ges, Palazzo dei Diamanti, Giulio d'Este, Oratorio, Palace, Schifanoia, San Francesco,	Ferrara	2	100	2.96	0.536
Renaissance, Architecture	Baroque, Music, Bugnato, Garofalo, Mantegna, Gonzaga, Sigismondo d'Este, Abstract Expressionism	Ferrara	3	58	1.30	0.404
Medieval, Castle	Medieval, Bacchanalia, Castle, Chapel, Court, Dawn, Dungeons, Lions, Loggia, Tower,	Ferrara	2	100	2.96	0.596
Medieval, Castle	Medieval, Castle, Bacchus, Calvinism, Diana, Endymnion, Evangelists, Pan	Ferrara	3	43	1.12	0.288

measured Average Node Connectivity, Focused Information Centrality and Focused Current Flow Betweenness Centrality. Note that the proposed semantic expansion achieves better precision results in presence of input terms linked to specific concepts related to the geographic location of the video. Considering, for example, the input user preferences "Sport, Cars" and the location "Modena", we note the expansion includes several related terms strongly linked to the city such as "Ferrari", "Maserati" etc. On the other hand the proposed solution obtains less accurate expansion in presence of term that are not strictly related to a specific location: for example the user preference "Fashion, Clothing, shopping" and the location "Modena". Furthermore, the Table presents the comparison results using a node expansion of two and three hops (see Section III-B). In most of the cases the two hops expansion achieves a better performance in both precision and centrality metrics. For instance, in the user preference "Architecture, Renaissance", location "Ferrara", the three hops expansion adds some terms less related, such as "Music" and "Abstract expressionism". Indeed, as the number of related nodes rears, the precision performance decreases from 91% to 72% and the value of centrality metric decreases by 23%, since the communities are enormously inflated and less focused.

Once the expanded user preferences is obtained we build a set of visual semantic classifiers. To analyze the effectiveness of our solution to detect user preferences into videos we annotate the full dataset (48 videos) according to a set of different user inputs. Table III shows the F1-Score results in the full dataset of four different strategies to grab online images for building visual classifiers:

TABLE III
COMPARISON OF SHOTS CLASSIFICATION F1-SCORE USING THE THREE AFOREMENTIONED APPROACHES

Method	F1-score
No Preprocessing (NP)	39.75
Semantic Expansion (SE)	42.91
Visual image filtering (VIF)	46.7
Semantic Expansion plus Metadata and visual image filtering (SE-VIF)	55.75

- *No preprocessing step:* we simply extract images from Visual Online Repositories (Flickr, Google) using the input user preferences.

- *Semantic expansion* we download images from Visual Online Repositories using the terms of the expanded preferences.

- *Visual image filtering:* we extract images from Visual Online Repositories using the input user preferences. Images are

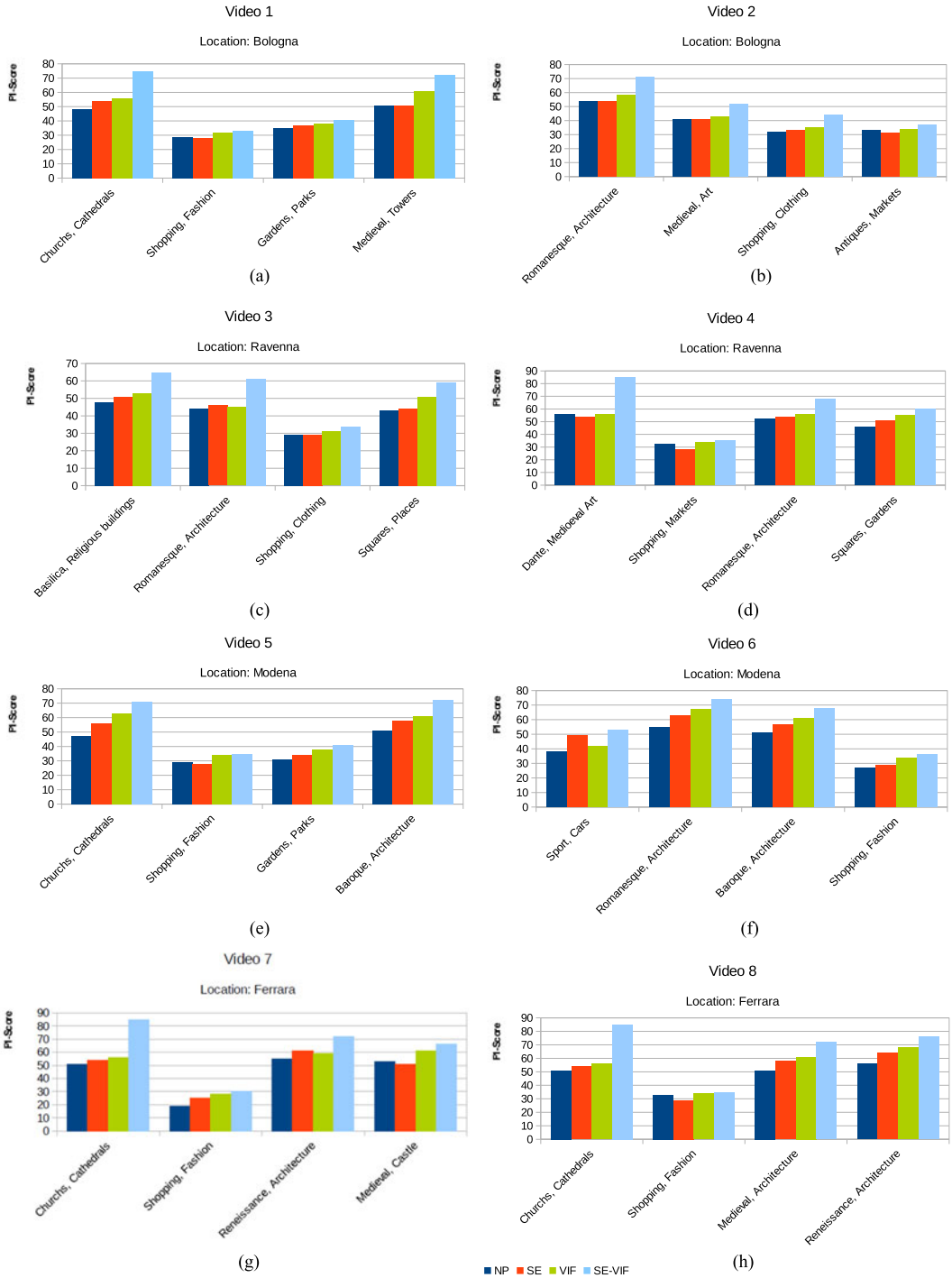

Fig. 7. For each video the classification comparison between classifiers trained with "NP"="No preprocessing", "SE"="Semantic Expansion", "VIF"="Visual Image Filtering", and "SE-VIF"="Semantic Expansion plus Metadata and Visual Image Filtering", as defined in Section IV-C is shown.

then filtered considering jointly visual and semantic features extracted by their metadata.

- *Semantic expansion plus metadata and visual image filtering (the proposed solution):* we extract images from Visual Online Repositories using the expanded set of the user preferences. Images are further filtered considering jointly visual and semantic features extracted by their metadata.

We note that both semantic expansion and visual image filtering strategies, taken alone, achieve better classification performance with respect to the "No preprocessing" baseline procedure. Moreover their combination, "Visual Preprocessing plus Semantic Preprocessing" strategy, further improves the user preference detection in the videos. In particular, Fig. 7 shows eight examples of classification comparison between the four approaches in term of F1-Score.

We observe that user preferences strongly related to cultural heritage sites of the city take a remarkable advantage. For instance, in Bologna, characterized by the presence of a large number of medieval, civil and religious, architecture buildings, user preferences such as "Medieval, Architecture" or "Church,

TABLE IV
COMPARISON RESULTS OF NARRATIVITY ASSESSMENT IN TERMS OF F1-SCORE

Attribute	Precision	Recall	F1-score
Uniform Baseline	23.95	38.80	29.62
DBSCAN Baseline	26.34	30.22	28.15
Our Narrative Approach	85.24	79.56	82.30

Cathedral" can be easily expanded semantically without introducing noisy terms . Therefore, the extracted images from online repositories are more coherent to the specific visit (Fig. 7 - videos: 1, 2). Similarly, this effect can be noted in the other art cites; for example see in Ferrara and Ravenna (Fig. 7 - videos: 3, 4 and 7, 8), respectively with "Renaissance, Architecture", "Medieval, Castle" and "Dante, Medieval Architecture" (Video 4, 7, 8). On the other hand, for example, in the "shopping" case the expansion procedure alone introduces hypernyms terms that decreases its specificity. However, the visual filtering mitigates this effect, since it performs a clustering based on visual and semantic features and removes outliers items. It is also interesting to notice that for the user preference "Sport, Cars" in location Modena (Video 6), "Semantic Expansion" alone remarkably improves the user preference detection in the videos, while the "Visual Image Filtering" alone reduces this improvement. This is presumably due to the fact that semantic expansion introduces highly specifically related terms, while visual filtering on images extracted without semantic expansion suffers from the large heterogeneity introduced by showroom sports cars images not embedded in real environments.

D. Experiments on Narrativity

To evaluate the proposed narrative assessment strategy we define two Baselines. The first baseline (called "Uniform") is obtained with a uniform sampling, i.e., given a video of length L, a desired length l and a minimum number of frames f for a ego-shot, an uniform selection of frame sequence at equal temporal length is made. For each selected frame the narrativity score is set to 1, otherwise 0. The second baseline (called "DBSCAN") is obtained using the DBScan agglomerative clustering over dense layer pre-trained $C3D$ net [48] features extracted from ego-shots, identified by using behavior pattern analysis. In particular, this network is trained from scratch to extract a visual feature by using a subset of 2000 shots of our dataset. Given a desired length l, we automatically fix the number of ego-shots to arrange the final video. We choose, first, ego-shots belonging to larger clusters (in particular the nearest to the pseudo-centroids). As before, we assign score equal to one to the selected shots (zero otherwise).

To analyze our methodology, a subset of videos is annotated identifying ego-shots that are important milestone in the narrative flow. Table IV presents experimental results in terms of Precision, recall and F1-Score.

As can be seen our method outperforms largely both the baselines. Uniform baseline performs indeed better than DBSCAN baseline, presumably due to the chronological smoothness naturally achieved in uniform approach, while DBSCAN makes no assumption on temporal flowing of the scenes. In particular, it emerges that a relevant fraction of uninformative shots, containing accidental captures of walls, cobbled pavings, asphalt roads, grids yards, crowds, traffic and busy streets, is filtered by means of informativeness assessment; lower but significant was the recall for highly informative shots.

Colorfulness classification helped to filter a large number of visually dull scenes, as the ones in which the viewer was staring at phone, timetables, road signs, walls.

E. Experiments on Summarization

Finally, to evaluate the effectiveness of our solution to obtain a summarized videos based on user preferences, we perform a "blind taste test" on a group of twelve users of different sexes and ages (15–40 years old). For each video a participant has to choose which summary best meets user preferences among our proposed solution, from now on addressed as "User Preferences" (UP), and four different approaches.

The four competitors consist of the upper-cited "Uniform" Baseline and "DBSCAN" Baseline (described in Section IV-D), an approach that we denominate "Different user Preferences" (DUP), and an approach denominated "Submodular Mixtures of Objectives". "Different User Preferences" (DUP) baseline is obtained by using our strategy, UP, but with different input terms, while "Submodular Mixtures of Objectives" (SMO) is the recent state-of-the-art-approaches proposed in [23].

We first show to the test group a browsable sped-up version of the entire original videos. Afterwards, for each original video, we show them the five summaries. The presentation order of the summaries is randomly produced for each original video. Therefore, the order of summaries changes every time and remains unknown to the test users.

After viewing all of them, each participant is asked to score summaries w.r.t. representativeness of user preference ("Fidelity to User Preference") and narrativity and temporal smoothness ("Chronological Visual Smoothness"), assigning an exclusive rate between 0 and 4 for each assessed attribute. As can be seen in Fig. 8, this test shows that the majority of the comparisons assigns a higher score to summaries obtained with our approach in both evaluation criteria with respect to "Submodular Mixture of Objectives", "Different user Preferences" and the other two baselines. It is worth noticing that "Submodular Mixture of Objectives" obtains the highest score among the rest competitors. In particular, its score is close to "User Preferences" approach for "Chronological Visual Smoothness", which is a consequence of its focus on temporal coherence and uniformity. Instead, "Submodular Mixture of Objectives" score is significantly lower for "Fidelity to User Preferences", which was expected since by design "User Preferences" approach is also directly focused on input query. It is also interesting to notice that "Uniform Baseline" performs generally better than DBscan, as it renders more smoothly the chronological storyline. This result is more marked in Narrativity assessment. In fact "DBscan", in most of the cases, creates a very jumpy and discontinuous summary, as lacking temporal smoothness. It is also interesting to notice that

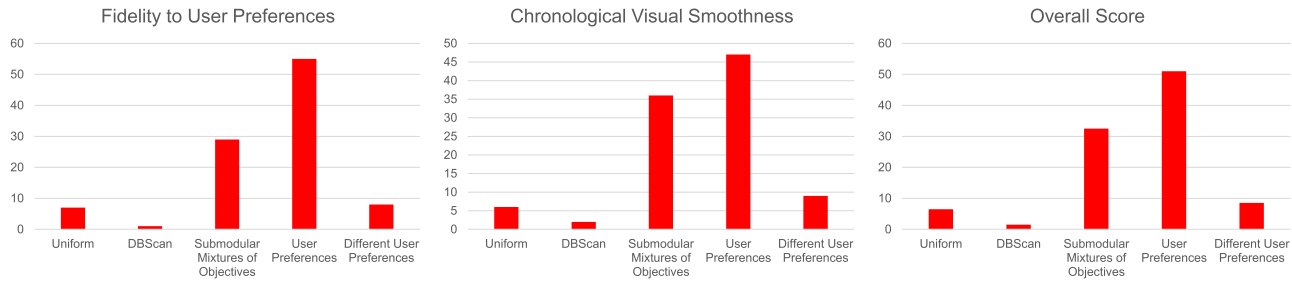

Fig. 8. "Blind taste test" user evaluation. It synthesizes the focus, narrativity, and overall scores obtained by the five compared methods: "Uniform", "DBScan", "Submodular Mixture of Objectives" [23], our approach labeled "User Preferences", and the same approach with different query input, labeled "Different User Preferences".

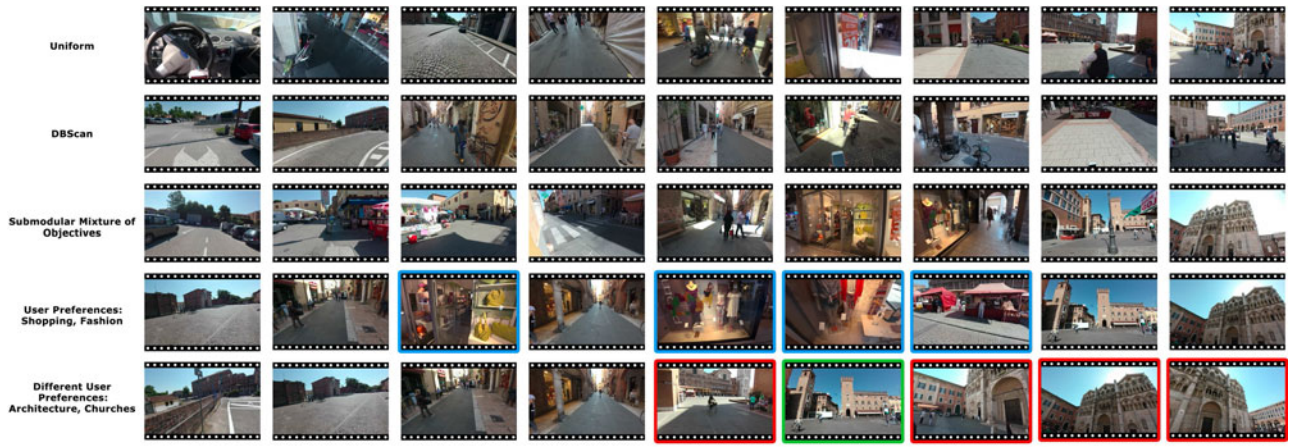

Fig. 9. Qualitative comparison of "Uniform", "DBScan", "Submodular Mixture of Objectives" [23], our approach labeled "User Preferences", and the same approach with different query input, labeled "Different User Preferences".

"Different user Preferences" approach performs also generally better than the two baselines. This is presumably due to the fact that, even if it is less or not at all focused on the topics of interest for the user, it provides both temporal smoothness, thanks to the narrativity component, and shakiness reduction due to the camera wearer's behavior assessment.

Finally, we present a qualitative evaluation of the produced summaries for each of the discussed methods. Due to the limited space, in Fig. 9 we show just an example of nine representative keyframes of resultant summaries from the same video captured in Ferrara. "Uniform" and "DBScan" Baselines generally extract relevant amounts of uninformative, low quality ego-shots and their recall is low on recovering user preference events. Nevertheless, while uniform baseline preserves temporal smoothness in the narrative flow of the original stream, DBScan often fails because it tends to focus on long visually homogeneous skims despite their temporal occurrence in the flow. "Submodular Mixture of Objectives" preserves a balanced trade-off between informativeness and temporal smoothness. "User Preferences" obtains also a significantly higher "Fidelity to User Preferences" score, being more focused on user preferences. For instance, Fig. 9 shows that User Preferences approach, focused on "Shopping, Fashion", selects a largest number of ego-shots capturing Shops and Markets, closely related to the input query, than "Submodular Mixture of Objectives" approach, preserving narrativity.

V. Conclusion

We proposed a new method for personalized video summarization in Cultural Heritage Domain where we formulated the problem in terms of behavior, semantic relatedness to the expressed user preferences and narrativity. Behavior is learned from visual apparent motion features and real motion features measured by GPS sensors. Relatedness to user preference is assessed by means of on-the-fly created semantic classifiers. Narrativity is measured according to shots importance and relevance assessed through a Personalized Page Rank approach, with learned model parameters. The experiments show interesting results since summaries produced with our method have been voted as the most relevant in terms of fidelity to user preferences and chronological visual smoothness.

References

[1] Z. Lu and K. Grauman, "Story-driven summarization for egocentric video," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2013, pp. 2714–2721.

[2] Y. Poleg, T. Halperin, C. Arora, and S. Peleg, "Egosampling: Fast-forward and stereo for egocentric videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4768–4776.

[3] Y. Poleg, A. Ephrat, S. Peleg, and C. Arora, "Compact cnn for indexing egocentric videos," in *Proc. IEEE Winter Conf. Appl. Comput. Vis.*, 2016, pp. 1–9.

[4] Y. J. Lee, J. Ghosh, and K. Grauman, "Discovering important people and objects for egocentric video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2012, pp. 1346–1353.

[5] M. Gygli, H. Grabner, R. Hayko, and L. Van Gool, "Creating summaries from user videos," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 505–520.
[6] B. Zhao and E. P. Xing, "Quasi real-time summarization for consumer videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2014, pp. 2513–2520.
[7] H. W. Ng, Y. Sawahata, and K. Aizawa, "Summarization of wearable videos using support vector machine," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2002, vol. 1, pp. 325–328.
[8] J. Xu *et al.*, "Gaze-enabled egocentric video summarization via constrained submodular maximization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 2235–2244.
[9] S. Swetha, A. Mishra, G. M. Hegde, and C. Jawahar, "Efficient object annotation for surveillance and automotive applications," in *Proc. IEEE Winter Appl. Comput. Vis. Workshops*, Mar. 2016, pp. 1–6.
[10] T. Shu, D. Xie, B. Rothrock, S. Todorovic, and S.-C. Zhu, "Joint inference of groups, events and human roles in aerial videos," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 4576–4584.
[11] B. Li and M. I. Sezan, "Event detection and summarization in sports video," in *Proc. IEEE Workshop Content-Based Access Image Video Libraries*, Dec. 2001, pp. 132–138.
[12] B. Li, H. Pan, and I. Sezan, "A general framework for sports video summarization with its application to soccer," in *Proc. IEEE Int. Conf. Acoust. Speech, Signal Process.*, Apr. 2003, vol. 3, pp. III-169–III-172.
[13] F. Chen and C. De Vleeschouwer, "Formulating team-sport video summarization as a resource allocation problem," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 21, no. 2, pp. 193–205, Feb. 2011.
[14] Z. Xiong, R. Radhakrishnan, A. Divakaran, Y. Rui, and T. S. Huang, "A unified framework for video summarization, browsing & retrieval: with applications to consumer and surveillance video," in *Image and Video Processing Handbook*. Orlando, FL, USA: Academic, 2006.
[15] Z. Ji, Y. Su, R. Qian, and J. Ma, "Surveillance video summarization based on moving object detection and trajectory extraction," in *Proc. 2nd Int. Conf. Signal Process. Syst.*, 2010, pp. V2-250–V2-253.
[16] Y. Peng and C.-W. Ngo, "Clip-based similarity measure for query-dependent clip retrieval and video summarization," *IEEE Trans. Circuits Syst. Video Technol.*, vol. 16, no. 5, pp. 612–627, May 2006.
[17] M. Wang *et al.*, "Event driven web video summarization by tag localization and key-shot identification," *IEEE Trans. Multimedia*, vol. 14, no. 4, pp. 975–985, Aug. 2012.
[18] M. Sun, A. Farhadi, B. Taskar, and S. Seitz, "Salient montages from unconstrained videos," in *Proc. 13th Eur. Conf. Comput. Vis.*, 2014, pp. 472–488.
[19] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Video summarization with long short-term memory," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 766–782.
[20] K. Zhang, W.-L. Chao, F. Sha, and K. Grauman, "Summary transfer: Exemplar-based subset selection for video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 1059–1067.
[21] A. G. del Molino, C. Tan, J. H. Lim, and A. H. Tan, "Summarization of egocentric videos: A comprehensive survey," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 65–76, Feb. 2017.
[22] M. Bolanos, M. Dimiccoli, and P. Radeva, "Towards storytelling from visual lifelogging: An overview," *IEEE Trans. Human-Mach. Syst.*, vol. 47, no. 1, pp. 77–90, 2017.
[23] M. Gygli, H. Grabner, and L. Van Gool, "Video summarization by learning submodular mixtures of objectives," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2015, pp. 3090–3098.
[24] A. Lidon *et al.*, "Semantic summarization of egocentric photo stream events," *CoRR*, 2015. [Online]. Available: http://arxiv.org/abs/1511.00438
[25] N. Joshi, W. Kienzle, M. Toelle, M. Uyttendaele, and M. F. Cohen, "Real-time hyperlapse creation via optimal frame selection," *ACM Trans. Graph.*, vol. 34, no. 4, 2015, Art. no. 63.
[26] M. Okamoto and K. Yanai, "Summarization of egocentric moving videos for generating walking route guidance," in *Proc. Pacific-Rim Symp. Image Video Technol.*, 2013, pp. 431–442.
[27] Y.-L. Lin, V. I. Morariu, and W. Hsu, "Summarizing while recording: Context-based highlight detection for egocentric videos," in *Proc. IEEE Int. Conf. Comput. Vis. Workshop*, Dec. 2015, pp. 443–451.
[28] T. Yao, T. Mei, and Y. Rui, "Highlight detection with pairwise deep ranking for first-person video summarization," in *Proc. IEEE Conf. Comput. Vis. Pattern Recog.*, Jun. 2016, pp. 982–990.
[29] H. W. Ng, Y. Sawahata, and K. Aizawa, "Summarization of wearable videos using support vector machine," in *Proc. IEEE Int. Conf. Multimedia Expo*, Aug. 2002, vol. 1, pp. 325–328.
[30] A. Sharghi, B. Gong, and M. Shah, "Query-focused extractive video summarization," in *Proc. 14th Eur. Conf. Comput. Vis.*, 2016, pp. 3–19.
[31] P. Varini, G. Serra, and R. Cucchiara, "Presonalized egocentric video summarization," in *Proc. 5th ACM Int. Conf. Multimedia Retrieval*, 2015, pp. 539–542.
[32] J. M. Henderson, "Regarding scenes," *Curr. Directions Psychological Sci.*, vol. 16, no. 4, pp. 219–222, 2007.
[33] G. Farnebäck, "Two-frame motion estimation based on polynomial expansion," in *Proc. 13th Scandinavian Conf. Image Anal.*, 2003, pp. 363–370.
[34] F. Crété-Roffet *et al.*, "The blur effect: Perception and estimation with a new no-reference perceptual blur metric," *Proc. SPIE*, vol. 6492, 2007, Art. no. 64920I.
[35] N. Cristianini and J. Shawe-Taylor, *An Introduction to Support Vector Machines and Other Kernel-Based Learning Methods*. Cambridge, U.K.: Cambridge Univ. Press, 2000.
[36] S. Auer *et al.*, "Dbpedia: A nucleus for a web of open data," in *Proc. 6th Int. Semantic Web 2nd Asian Conf. Asian Semantic Web Conf.*, 2007, pp. 722–735.
[37] E. W. Dijkstra, "A note on two problems in connexion with graphs," *Numerische Mathematik*, vol. 1, pp. 269–271, 1959.
[38] M. E. Newman and M. Girvan, "Finding and evaluating community structure in networks," *Phys. Rev. E*, vol. 69, no. 2, 2004, Art. no. 026113.
[39] U. Brandes and D. Fleischer, "Centrality measures based on current flow," in *Proc. 22nd Annu. Conf. Theoretical Aspects Comput. Sci.*, 2005, pp. 533–544.
[40] L. Beineke, O. Oellermann, and R. Pippert, "The average connectivity of a graph," *Discrete Math.*, vol. 3, pp. 31–45, 2002.
[41] I. Hulpus, C. Hayes, M. Karnstedt, and D. Greene, "Unsupervised graph-based topic labelling using dbpedia," in *Proc. 6th ACM Int. Conf. Web Search Data Mining*, 2013, pp. 465–474.
[42] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," arXiv preprint arXiv:1409.1556, 2014.
[43] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *CoRR*, 2013. [Online]. Available: http://arxiv.org/abs/1301.3781
[44] M. Ester, H.-P. Kriegel, J. Sander, and X. Xu, "A density-based algorithm for discovering clusters in large spatial databases with noise," in *Proc. 2nd Int. Conf. Knowl. Discovery Data Mining*, 1996, pp. 226–231.
[45] S. D. Kamvar, T. H. Haveliwala, C. D. Manning, and G. H. Golub, "Extrapolation methods for accelerating pagerank computations," in *Proc. 12th Int. Conf. World Wide Web*, 2003, pp. 261–270.
[46] A. N. Langville, C. D. Meyer, and P. FernÁndez, "Googles pagerank and beyond: The science of search engine rankings," *Math. Intelligencer*, vol. 30, no. 1, pp. 68–69, 2008.
[47] D. A. Schult and P. Swart, "Exploring network structure, dynamics, and function using networkx," in *Proc. 7th Python Sci. Conf.*, 2008, pp. 11–15.
[48] D. Tran, L. Bourdev, R. Fergus, L. Torresani, and M. Paluri, "Learning spatiotemporal features with 3d convolutional networks," in *Proc. IEEE Int. Conf. Comput. Vis.*, Dec. 2015.
[49] X. Lu, Z. Lin, H. Jin, J. Yang, and J. Z. Wang, "Rapid: Rating pictorial aesthetics using deep learning," in *Proc. ACM Int. Conf. Multimedia*, 2014, pp. 457–466.
[50] R. Collobert and J. Weston, "A unified architecture for natural language processing: Deep neural networks with multitask learning," in *Proc. 25th Int. Conf. Mach. Learn.*, 2008, pp. 160–167.

Patrizia Varini received the Laurea degree in physics from the University of Bologna, Bologna, Italy, in 1993, and the Ph.D. degree in computer engineering, multimedia, and telecommunication from the University of Modena and Reggio Emilia, Modena, Italy, in 2017.

Her research interests include semantic multimedia analysis, indexing and retrieval, social media and big data analytics, knowledge management, and personalization for multimedia applications.

Giuseppe Serra is currently an Assistant Professor with the University of Udine, Udine, Italy. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, USA, and at Telecom ParisTech/ENST, Paris, France, in 2006 and 2010, respectively. He has authored or coauthored more than 40 publications in scientific journals and international conference proceedings. His research interests include egocentric vision, and image and video analysis.

Prof. Serra has been an Associate Editor for the IEEE TRANSACTIONS ON HUMAN-MACHINE SYSTEMS since 2017. He was a Technical Program Committee member of several workshops and conferences. He regularly serves as a Reviewer for international conferences and journals such as CVPR and ACM Multimedia.

Rita Cucchiara (M'99) is currently a Full Professor with the University of Modena and Reggio Emilia, Modena, Italy. She is the Director of the Research Center Softech-ICT and heads the Imagelab Laboratory at the University of Modena and Reggio Emilia. She has been a coordinator of several projects in computer vision and pattern recognition, and in particular on video surveillance, human behavior analysis, and video understanding.

Prof. Cucchiara is a Member of the ACM, a Member of the IEEE Computer Society, and a Fellow of the IAPR. She is the President of the Gruppo Italiano Ricercatori in pattern recognition (affiliated with IAPR), and a Member of the Advisory Board of the Computer Vision Foundation.