# Web Mining & Computer Vision: New Partners for Object-based Activity Recognition

Daniele Riboni Marta Murtas
Department of Mathematics and Computer Science, University of Cagliari
via Ospedale 72, I-09122 Cagliari, Italy
Email: riboni@unica.it, murtasmarta@gmail.com

***Abstract*—In several domains, including healthcare and home automation, it is important to unobtrusively monitor the activities of daily living (ADLs) executed by people at home. A popular approach consists in the use of sensors attached to everyday objects to capture user interaction, and ADL models to recognize the current activity based on the temporal sequence of used objects. However, both knowledge-based and data-driven approaches to object-based ADL recognition have different issues that limit their applicability in real-world deployments. Hence, in this paper, we pursue an alternative approach, which consists in mining ADL models from the Web. Existing attempts in this sense are mainly based on Web page mining and lexical analysis. One issue with those attempts relies on the high level of noise found in the textual content of Web pages. In order to overcome that issue, our intuition is that pictures illustrating the execution of a given activity offer much more compact and expressive information than the textual content of a Web page regarding the same activity. Hence, we present a novel method to couple Web mining and computer vision for automatically extracting ADL models from visual items. Our method relies on Web image search engines to select the most relevant pictures for each considered activity. We use off-the-shelf computer vision APIs and a lexical database to extract the key objects appearing in those pictures. We introduce a probabilistic technique to measure the relevance among activities and objects. Through experiments with a large dataset of real-world ADLs, we show that our method significantly improves the existing approach.**

## I. Introduction

Activity recognition is a key requirement in several pervasive computing domains [1], including smart home automation, e-health, gaming, manufacturing, pervasive advertising, and smart cities. Currently, the most popular approach to activity recognition consists in the use of supervised learning methods applied to datasets of activities and sensor data [2]. Supervised learning proves to be effective in recognizing activities characterized by specific postures or motions, such as physical activities. However, the actual applicability of the supervised approach to complex ADLs (e.g., cleaning, dressing, eating) is questionable, especially when infrequent or sporadic activities are taken into account (e.g., “wrap a gift”, “prepare a lunch box”). Indeed, acquiring large datasets of activities is expensive in terms of annotation costs [3]. Moreover, activity annotation by an external observer, by means of cameras or direct observation, violates user privacy.

For this reason, different researchers tried to devise unsupervised methods for recognizing ADLs based on sensor data. Unsupervised methods rely on a symbolic modeling of activities in terms of their constituting simpler actions. For instance, the temporal sequence of events “open medicine cabinet; take medicine box; put away medicine box; close medicine cabinet” characterizes the ADL “taking medicines”. A popular direction is to manually define those models through formal ontologies expressed in a description logics language [4]. However, manually defining comprehensive ADL ontologies is cumbersome. Moreover, the ontological approach is generally based on rigid activity definitions, that fall short in adapting to dynamic context conditions. An alternative approach relies on Web mining to automatically infer the activity model from Web pages regarding that activity. That approach is generally denoted as *Web-based activity mining*. In particular, different works tried to associate an activity with the objects related to its execution by mining the textual content of Web pages regarding that activity: by monitoring the use of those objects through sensors, it would be possible to automatically reconstruct the activity [5], [6], [7].

One limitation of current Web-based activity mining methods relies on the high level of noise found in the textual content of Web pages. To overcome that issue, our research starts from the intuition that images showing the execution of a given activity provide more concise and expressive information than the text of a Web page regarding that activity. For instance, consider the pictures illustrated in Figure 1, which have been chosen among the top-10 results of Google Images[1] search for the query “cleaning kitchen”. Those pictures include several objects (broom, rag, bucket, sink...) typically used to perform that activity. Moreover, they capture different ways to perform the same activity. Note that the pictures concentrate on the key objects regarding that activity, disregarding irrelevant ones. In this paper, we aim at taking the Web-based activity mining approach one step further, by exploiting Web image search engines and computer vision APIs. To our knowledge, this is the first work that investigates this direction. The main contributions of this work are the following:

- We illustrate a novel method to couple Web image search, computer vision functions, and lexical analysis to extract the most relevant objects with respect to activities of interest.
- We introduce a probabilistic weighting scheme to mea-

[1] https://images.google.com/

Fig. 1. Images related to “cleaning kitchen” retrieved by Google image search

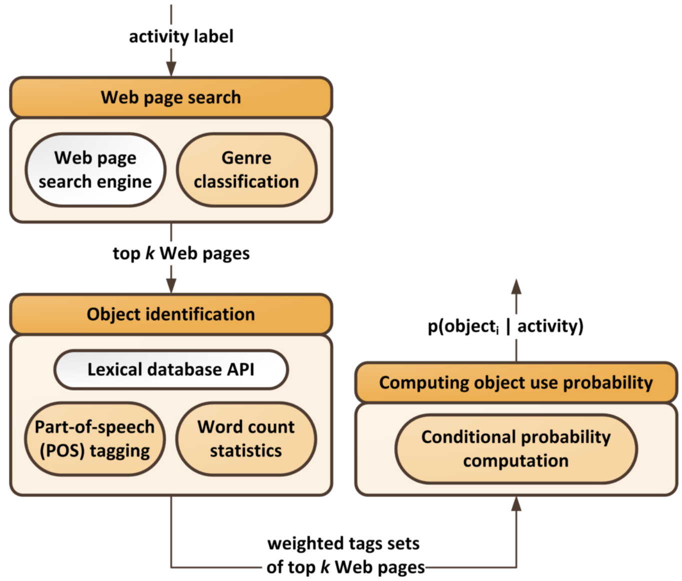



Fig. 2. Web-based activity mining method based on Web page text.

sure the relevance among activities and objects.

- We present the results of extensive experiments with a large dataset of real-world ADLs, and show that our method outperforms the existing Web-based activity mining method based on textual content.

The rest of the paper is structured as follows. Section II illustrates related work. Section III introduces our method. Section IV presents the experiments and discusses results. Section V concludes the paper.

## II. Related work

As explained in the introduction, the supervised learning approach to the recognition of complex ADLs has different limitations in terms of generality, applicability, costs, and privacy. Hence, different efforts have been made in the last years to devise unsupervised ADLs recognition methods.

In particular, several researchers designed activity ontologies expressed through description logic languages, possibly extended with rules, to model ADLs based on their constituting simpler actions [8], [9], [10]. Activity recognition is based on the observation of temporal sequences of sensor events that match the definition of actions defining a given activity. As explained before, that approach has limits in the rigidity of activity descriptions and burden of manually defining the ontological axioms.

A third approach, which aims at taking the best of automatic activity modeling and unsupervision, consists in mining activity models from the Web. A first attempt in this sense was due to Perkowitz et al. in [5] and refined in later works [11], [12]. The approach is illustrated in Fig. 2. The input is an activity label such as “cleaning kitchen”. The activity label is used as a query for a Web search engine, to find pages related to that activity. A *genre classification* module is used to select only those pages containing detailed description of that activity. Then, the textual content of the top $k$ pages is passed to an object identification module, which exploits a lexical database to extract the key objects related to the activity. In particular, part-of-speech (POS) tagging is used to prune terms that do not refer to objects. A statistical method is used to obtain the set of top-$j$ related objects; each object is weighted considering its frequency in the Web pages. Finally, for each object $o \in O$ ($O$ being the set of objects) with weight $w$, and each activity $a \in A$ ($A$ being the set of activities), the probability of using $o$ during $a$ is computed. The probability distribution $p(O|A)$ is used by a generative model to reconstruct the most probable sequence of activities given an observed sequence of used objects. That method is unsupervised, except for the genre classification module that relies on supervised machine learning.

A similar approach was used by Pentney et al. in [13], exploiting user-contributed common sense acquired by the Open Mind Indoor Common Sense project [14]. Gu et al. presented a different method to extract activity models from the text of Web pages [15]. In that work, object-use fingerprints are extracted in terms of *contrast patterns*, which describe statistically significant differences in object-use patterns among any couple of activities. A similar method was used by Palmes et al. in [6] for both activity recognition and segmentation. Ihianle et al. propose a method to mine the textual content of Web pages for identifying the most probable activity given a temporal sequence of used objects [7]. The most probable activities are inferred based on a combination of statistical and ontological reasoning.

Even though they adopt different techniques for inferring the relevance among objects and activities, to the best of our knowledge all existing Web-based activity mining methods rely on textual content only. Instead, in this work we investigate a different approach: exploiting the visual content of Web pages to derive the activity model.

### III. Methodology and Algorithms

In this section, we illustrate our methodology to extract correlations among objects and activities, as well as the algorithms to implement the method.

#### A. Methodology

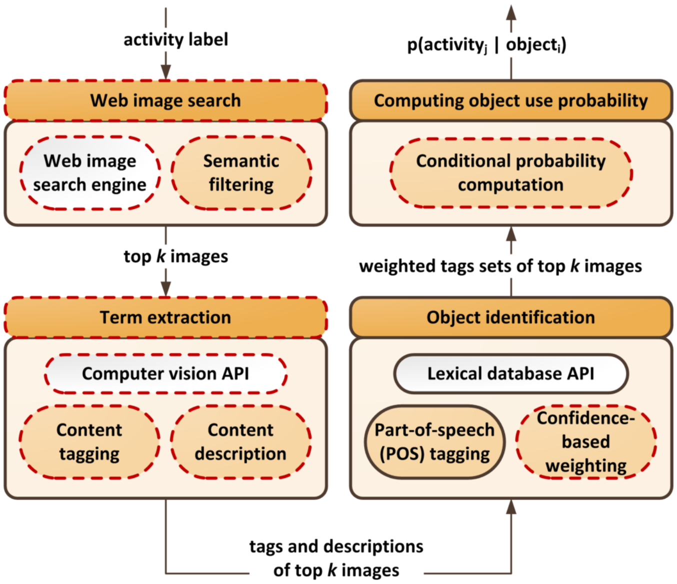



Fig. 3. Our Web-based activity mining method based on image search and computer vision. Modules highlighted in red (dotted line) are novel or different from the existing approach based on Web page text. Modules in white represent external APIs.

Our method is illustrated in Fig. 3. Given an activity label (e.g., "dish washing"), the first module (Web image search) queries a Web image search engine to find the top $k$ images that match the label and satisfy a semantic filtering directive. The goal of semantic filtering is to prune those images which are returned by the search engine, but are semantically unrelated to the activity. Semantic filtering relies on an analysis of the textual content of the Web page that contains the image.

The top $k$ images are given to the Term extraction module, which queries a Computer vision API to extract a description and tags of elements identified in the images. The description briefly summarizes the image content (e.g., "a person washing a cup"). Image tags refer to objects and other elements found in the image (e.g., dish, water, white). Each tag is associated to a confidence value, which represents the probability that the element actually appears in the image.

Tags and descriptions of the top $k$ images are passed to the Object identification module, which applies POS tagging to keep only those tags that are nouns denoting objects. POS tagging is applied also to the image description, in order to extract other tags denoting objects. Tags extracted from the image description are assigned confidence 1. For each tag found in an image, we compute a weight based on its confidence values.

The above procedure is executed for each activity $a$ belonging to the set $A$ of considered activities. Finally, weighted objects from the top $k$ images of each activity are used by the module for Computing object use probability, which computes the correlations among objects and activities. In particular, for each considered activity $a$, and each object $o$ found in at least one image, that module estimates $p(a|o)$; i.e., the conditional probability that the current activity is $a$ given that the used object is $o$.

#### B. Algorithms

The algorithm *ObjIdentification* implements the modules for Web image search, Term extraction and Object identification, shown in Fig. 3. It takes as input an activity label $a$ and the number $k$ of images to be used to identify the objects related to the execution of $a$. The output is the set of identified objects for each image, together with their respective weight. After initializing a set $imgs$ to the empty set, the algorithm queries a Web image search engine and downloads the top $k$ images that respond to the query '$a$'. Those images are added to $imgs$. For each image, a semantic filtering method is used to analyze the textual content of the Web page containing it, in order to check whether the image is actually related to the search query. In its simplest form, the method checks whether the Web page contains the text '$a$': if not, the image is removed from *imgs*. At the first step, if any image is removed, the size of $imgs$ is less than $k$. Hence, the algorithm queries the image Web search engine to download other $k$ images, and applies semantic filtering. This process is repeated until $imgs$ contains at least $k$ images. Then, the algorithm selects the top $k$ images according to the order of selection by the search engine.

For each image $img_i$, the algorithm queries a visual analysis API to get a set of tags referring to entities appearing in $img_i$, as well as a textual description of the image content. Each tag is associated to a confidence value, ranging from 0 to 1, which represents the probability of that entity to actually appear in $img_i$ according to the computer vision algorithm. For each tag, the algorithm queries a POS tagger to get its lexical category (object, plant, animal, etc.); those tags that do not refer to objects or artifacts are removed from the set of tags. Then, the POS engine is queried to extract additional terms from the image description. Terms not referring to objects or artifacts are discarded. For each remaining term, the algorithm checks whether it appears as the label of any tag of the image. If so, the confidence of that tag for the image is set to 1. Otherwise, a new tag with that label is created for that image, and its confidence value is set to 1. Finally, the set $T_a$ of weighted tags sets $tags_1, \ldots, tags_k$ obtained from the top $k$ images is returned.

The algorithm *ObjProb* implements the module for Computing object use probability shown in Fig. 3. It takes as input the set of considered activity labels $a_1, \ldots, a_n$ and the number $k$ of images per activity. The output is the set of conditional probabilities $p(a_j|o_i)$, for each activity $a_j$ and object $o_i$. At first, for each $a_j$, the algorithm executes the *ObjIdentification*$(a_j, k)$ algorithm to get the set $T_{a_j} = \{tags_1, \ldots, tags_k\}$ of weighted tags sets associated to $a_j$. For each $tag$ of each tags set $tags_i$, the algorithm assigns the tag's confidence $tag.conf$ to $p(tag.label, a_j, i)$. The latter is the probability of observing the object corresponding to the tag's label in the $i^{th}$ image of $a_j$. Then, for each activity $a_j$ and each object $o_i$, the algorithm computes the conditional probability $p(a_j|o_i)$ according to the following formula:

$$p(a_j|o_i) = \frac{\sum_{l} p(o_i, a_j, l)}{\sum_{m,l} p(o_i, a_m, l)}$$

Finally, the algorithm returns the conditional probability distribution $P(A|O)$, where $A$ is the set of activities and $O$ is the set of objects.

## IV. Experimental evaluation

In order to validate our approach, we performed extensive experiments with a real-world dataset of ADLs executed in a smart home. We compared our method with an existing Web-based activity mining method, which relies on Web page search and lexical analysis. To provide the possibility to replicate the experiments, we publish our code online[2].

### A. Experimental setup

In our experiments, we used the well-known dataset of Cook et al. [16], [17], named *CASAS*, which includes both interleaved and sequential ADLs executed in a smart-home by twenty-one subjects[3]. Sequential activities are pre-segmented, while interleaved activities are not. Since, in the real world, people perform activities in an interleaved fashion, we limit our attention to the recognition of interleaved ADLs. Sensors collected data about movement, presence in specific home locations, temperature, use of water, interaction with objects, doors, phone; 70 sensors were used in total. In our work, we considered only 24 out of 70 sensors. Indeed, the other sensors (mostly presence sensors) were not associated to the use of objects or furniture.

The dataset considers eight activities, whose labels are reported in Table I. The order and time taken to perform the activities were up to the subject. Activities were executed naturalistically by a single subject at a time. In the dataset, each sensor activation (e.g., "fridge opened", "cup moved") is labeled with the timestamp of the event, and with the current activity executed at that timestamp. Given the temporal sequence of sensor activations, the goal of the activity recognition system is to reconstruct the current activity at each activation.

### B. Activity mining using text and lexical analysis

In order to compare our method with existing Web-based activity mining ones, we have developed a method based on Web page search and lexical analysis similar to the one illustrated in Figure 2. The algorithms were implemented in Python. For each activity label, we downloaded the first $k$ Web pages found by Google Search. We fixed $k$ to 15, because we experimentally found that it was the optimal value. For each Web page, we saved the content of the ⟨body⟩ element. For each word in the body, we looked for its POS category using WordNet; i.e., a lexical database of English nouns, verbs, adjectives and adverbs [18]. We queried WordNet using its API called Natural Language Toolkit (NLTK)[4], which allows, among other things, to find the POS of words. We saved only those words that are objects or artifacts, and we computed the number of occurrences and the *weight* of every word. The weight is computed as the probability of that word to actually be a noun (# of senses of that word that are noun / total # of senses of that word). For each word we found its synonyms (they also have to be objects or artifacts), for making the technique as flexible as possible. Finally, we computed the "object-use probability" according to the algorithm *ObjProb* explained in Section III-B, but considering weighted tag sets extracted from Web page text instead of images. Note that, differently from the technique presented in [11], we compute the distribution of $p(\textit{activity}|\textit{object})$, since we use a discriminative activity recognition algorithm.

### C. Activity mining using images and computer vision

We implemented our algorithms, presented in Section III-B, in Python. Based on the activity label, we downloaded the top $k$ pictures from Google Images (they have to be photos of medium dimensions). Also in our case, $k$ was experimentally set to 15. For the sake of this paper, we did not implement the *semantic filtering* module; i.e., we kept all images irrespectively of the Web page textual content. Then, we used the Computer Vision APIs of Microsoft Cognitive Services[5] to get the set of *tags* identified in every image. Tags may be objects or other features describing the image content, such as "indoor", or "yellow". In order to retain only nouns referring to artifacts or objects, we used the NLTK APIs, applying the same method described in Section IV-B. Each tag returned by the computer vision APIs is associated to a confidence value in $(0; 1]$. For every activity $a$ and object $o$, we computed the conditional probability $p(a|o)$ according to algorithm *ObjProb*.

### D. Activity recognition method

In our experiments, given a temporal sequence of sensor events $\langle e_1, e_2, \ldots, e_n \rangle$ and their respective timestamps $\langle t_1, t_2, \ldots, t_n \rangle$, the objective of activity recognition is to reconstruct the current activity at each $t_i$. We denote by $O$ the set of objects whose use is monitored by the smart home sensors, and by $A$ the set of considered activities. Each event $e_j$ captures the interaction with an object $o(e_j) \in O$ at $t_j$ during the execution of an activity.

For the sake of this work, we devised a simple activity recognition method, which considers the conditional probabilities $p(A|O)$ and a fixed-length sliding window of the $n$ most recent events. In particular, for each timestamp $t_j$ ($j \geq n$) and for each activity $a \in A$, we compute the weight $w(a, t_j)$, which is the temporally smoothed product of the

[2]http://people.unica.it/danieleriboni/python_code/
[3]http://ailab.wsu.edu/casas/datasets/adlinterweave.zip
[4]http://www.nltk.org/
[5]https://www.microsoft.com/cognitive-services/en-us/computer-vision-api

conditional probability of $a$ being the current activity at $t_j$, $t_{j-1}, \ldots, t_{j-n+1}$. Formally:

$$w(a, t_j) = \prod_{k=j-n+1\ldots j} p(a|o(e_j)) \cdot c^{j-k},$$

where $c \in (0, 1]$ is the temporal smoothing factor, used to give more relative weight to the recent events. Given the weights computed for each activity at $t_j$, the predicted activity $pred(t_j)$ is the one that maximizes the weight; formally:

$$pred(t_j) = a \in A \; such \; that \, w(a, t_j) = \max_{a \in A} \{w(a, t_j)\}.$$

### E. Results and discussion

We evaluated the prediction's quality in terms of the standard measures of precision, recall and $F_1$ score; the latter is the harmonic mean of precision and recall. Since the methods are unsupervised, we did not use cross-validation. In the first experiments, we set the temporal smoothing factor $c$ to 1; i.e., no temporal smoothing. We varied the size $n$ of the sliding window from 1 to 10. Results are shown in Fig. 4. The best results are obtained using relatively small values of $n$. With no temporal smoothing, our image-based method significantly outperformed the text-based method ($F_1 = 0.6876$ vs $F_1 = 0.6018$).

Then, for each method, we fixed the size $n$ of the sliding window, and tried different values of temporal smoothing factor $c$ from 0.1 to 1. Fig. 5 reports the results. With the text-based method, we obtained the best results fixing $n = 2$. The highest $F_1$ score 0.5998 was obtained with $c = 0.8$. With our image-based method, the highest $F_1$ score 0.6988 was obtained with $n = 3$ and $c = 0.5$. In general, with both methods, the influence of temporal smoothing was limited.

Table I reports the average precision, recall and $F_1$ score for each activity. We found some activities (especially, "watering plants", "answering the phone", and "hoovering") hard to recognize, mainly due to the lack of sufficient information about objects usage in the dataset for those activities, which determined low recall values. Precision was reasonably good with most activities. However, the precision of "watering plants" was low. Indeed, as shown by the confusion matrix in Table II, that activity was frequently confused with "preparing soup". Those errors happened because both activities involve the use of the same or similar tools (e.g., sinks and water containers). We claim that this is an intrinsic limit of object based-activity recognition; it is not due to the specific method used to define object-based activity models. This problem can be mitigated by considering additional smart objects that can better characterize the activities. With most activities, the values of recall were lower than those of precision. The reason is that some objects had weight 0 for every activity; hence, activities in which they were used were assigned to class *null* (last column of the confusion matrix).

A comparison with other activity recognition techniques using the same dataset indicates that our results are promising. The Hidden Markov Model method used in [16] achieved average $F_1 = 0.700$; our method achieved essentially the same score, having the advantage of being unsupervised (no training set must be acquired). The unsupervised method presented in [19], based on a hybrid combination of ontological and probabilistic reasoning, achieved higher accuracy; i.e., average $F_1 = 0.781$. However, that method adopts an advanced activity recognition technique, while the method used in this paper was rather simplistic. Moreover, that hybrid method strongly relies on manual activity modeling through ontology engineering, while our method has the advantage of being fully automatic.

TABLE I
AVERAGE PRECISION, RECALL AND F1 SCORE FOR EACH ACTIVITY. WEB-BASED ACTIVITY MINING USING IMAGES AND COMPUTER VISION. $n = 3$, $c = 0.5$.

| **activity** | **label** | **precision** | **recall** | **F1 score** |
|---|---|---|---|---|
| $a_1$ | fill medicine cabinet | 0.9334 | 0.6132 | 0.7401 |
| $a_2$ | watch television | 0.776 | 0.9241 | 0.8436 |
| $a_3$ | watering plants | 0.3418 | 0.3799 | 0.3598 |
| $a_4$ | answering the phone | 0.7907 | 0.3506 | 0.4858 |
| $a_5$ | diy birthday card | 0.7399 | 0.9464 | 0.8305 |
| $a_6$ | prepare soup | 0.7223 | 0.6224 | 0.6686 |
| $a_7$ | hoovering | 0.77 | 0.237 | 0.3624 |
| $a_8$ | choosing outfit | 0.976 | 0.5365 | 0.6924 |
| all | all activities | 0.729 | 0.6709 | 0.6988 |

TABLE II
CONFUSION MATRIX FOR WEB-BASED ACTIVITY MINING USING IMAGES AND COMPUTER VISION. $n = 3$, $c = 0.5$.

| **classified as** → | $a_1$ | $a_2$ | $a_3$ | $a_4$ | $a_5$ | $a_6$ | $a_7$ | $a_8$ | *null* |
|---|---|---|---|---|---|---|---|---|---|
| $a_1$ | **84** | 7 | 5 | 0 | 13 | 9 | 3 | 0 | 16 |
| $a_2$ | 0 | **426** | 0 | 0 | 23 | 0 | 0 | 0 | 12 |
| $a_3$ | 0 | 13 | **68** | 2 | 2 | 80 | 0 | 0 | 14 |
| $a_4$ | 0 | 3 | 0 | **34** | 58 | 0 | 0 | 2 | 0 |
| $a_5$ | 1 | 5 | 2 | 0 | **529** | 0 | 0 | 0 | 22 |
| $a_6$ | 5 | 5 | 25 | 5 | 21 | **234** | 12 | 0 | 69 |
| $a_7$ | 0 | 65 | 96 | 0 | 62 | 0 | **77** | 0 | 25 |
| $a_8$ | 0 | 25 | 3 | 2 | 7 | 1 | 8 | **81** | 24 |

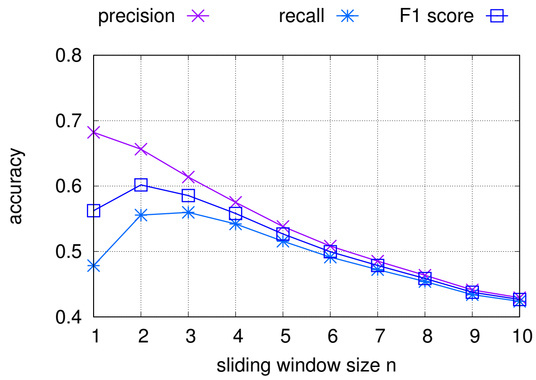
## V. CONCLUSION AND FUTURE WORK

In this paper, we introduced a new approach to Web-based activity mining, which relies on visual information, instead of textual content. We presented a technique for extracting relevant activity images from the Web, identifying key objects through computer vision functions, and computing activity-object relevance weights. A detailed experimental comparison with related works shows the potential of our approach.

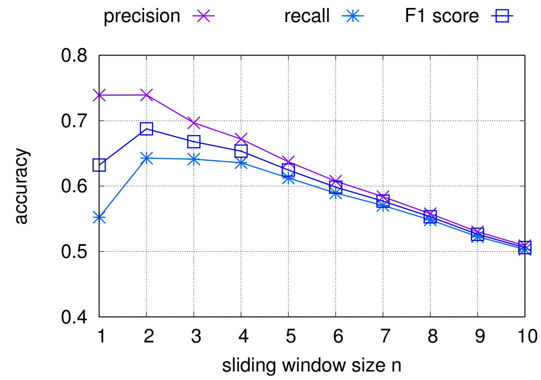
This work can be extended in several directions. The technique to select relevant images could be improved by analyzing the textual content of the Web page in which they appear. While we used general-purpose computer vision APIs, object recognition could be improved adopting specific methods to recognize human-object interaction. Last but not least, temporal activity information could be extracted by mining activity data from videos instead of images.

## ACKNOWLEDGMENT

This work was partially supported by the "DomuSafe" project, funded by Sardinia regional government (CRP 69, L.R. 7 agosto 2007, n.7).

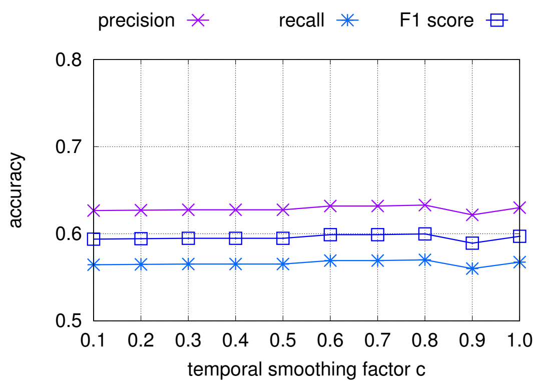



(a) Web-based activity mining using text and lexical analysis.

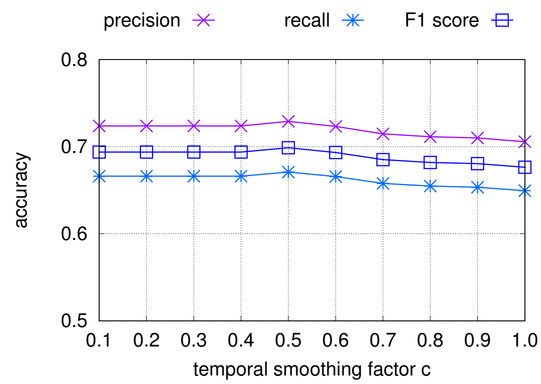



(b) Web-based activity mining using images and computer vision.

Fig. 4. Results according to the size $n$ of the sliding windows. No temporal smoothing ($c = 1$).




(a) Web-based activity mining using text and lexical analysis. The sliding window size is fixed to $n = 2$.




(b) Web-based activity mining using images and computer vision. The sliding window size is fixed to $n = 3$.

Fig. 5. Results according to the temporal smoothing factor $c$. The size of the sliding window is fixed.

## REFERENCES


[1] N. Davies, D. P. Siewiorek, and R. Sukthankar, "Activity-based computing," *IEEE Pervasive Computing*, vol. 7, no. 2, pp. 20–21, 2008.

[2] L. Bao and S. S. Intille, "Activity recognition from user-annotated acceleration data," in *Proceedings of Pervasive Conference*. Springer, 2004, pp. 1–17.

[3] A. Calatroni, D. Roggen, and G. Tröster, "Collection and curation of a large reference dataset for activity recognition," in *Proceedings of Systems, Man, and Cybernetics Conference*. IEEE, 2011, pp. 30–35.

[4] D. Riboni and C. Bettini, "OWL 2 modeling and reasoning with complex human activities," *Pervasive and Mobile Computing*, vol. 7, no. 3, pp. 379–395, 2011.

[5] M. Perkowitz, M. Philipose, K. P. Fishkin, and D. J. Patterson, "Mining models of human activities from the web," in *Proceedings of WWW Conference*. ACM, 2004, pp. 573–582.

[6] P. P. Palmes, H. K. Pung, T. Gu, W. Xue, and S. Chen, "Object relevance weight pattern mining for activity recognition and segmentation," *Pervasive and Mobile Computing*, vol. 6, no. 1, pp. 43–57, 2010.

[7] I. K. Ihianle, U. Naeem, A. H. Tawil, and M. A. Azam, "Recognizing activities of daily living from patterns and extraction of web knowledge," in *UbiComp Adjunct Proceedings*. ACM, 2016, pp. 1255–1262.

[8] L. Chen and C. Nugent, "Ontology-based activity recognition in intelligent pervasive environments," *International Journal of Web Information Systems*, vol. 5, no. 4, pp. 410–430, 2009.

[9] D. Riboni, L. Pareschi, L. Radaelli, and C. Bettini, "Is ontology-based activity recognition really effective?" in *Proceedings of the Ninth Annual IEEE International Conference on Pervasive Computing and Communications Workshops*. IEEE, 2011, pp. 427–431.

[10] R. Helaoui, D. Riboni, and H. Stuckenschmidt, "A probabilistic ontological framework for the recognition of multilevel human activities," in *Proceedings of ACM UbiComp*. ACM, 2013, pp. 345–354.

[11] D. Wyatt, M. Philipose, and T. Choudhury, "Unsupervised activity recognition using automatically mined common sense," in *Proceedings AAAI*. AAAI Press / The MIT Press, 2005, pp. 21–27.

[12] E. M. Tapia, T. Choudhury, and M. Philipose, "Building reliable activity models using hierarchical shrinkage and mined ontology," in *Proceedings of Pervasive*, ser. LNCS, vol. 3968. Springer, 2006, pp. 17–32.

[13] W. Pentney, A. Popescu, S. Wang, H. A. Kautz, and M. Philipose, "Sensor-based understanding of daily life via large-scale use of common sense," in *Proceedings of AAAI*. AAAI Press, 2006, pp. 906–912.

[14] R. Gupta and M. J. Kochenderfer, "Common sense data acquisition for indoor mobile robots," in *Proceedings of AAAI*. AAAI Press / The MIT Press, 2004, pp. 605–610.

[15] T. Gu, S. Chen, X. Tao, and J. Lu, "An unsupervised approach to activity recognition and segmentation based on object-use fingerprints," *Data Knowl. Eng.*, vol. 69, no. 6, pp. 533–544, 2010.

[16] G. Singla, D. J. Cook, and M. Schmitter-Edgecombe, "Tracking activities in complex settings using smart environment technologies," *Int. J. Biosci. Psychiatr. Technol.*, vol. 1, no. 1, pp. 25–35, 2009.

[17] D. J. Cook, A. S. Crandall, B. L. Thomas, and N. C. Krishnan, "CASAS: A smart home in a box," *Computer*, vol. 46, no. 7, pp. 62–69, 2013.

[18] G. A. Miller, "Wordnet: A lexical database for english," *Commun. ACM*, vol. 38, no. 11, pp. 39–41, 1995.

[19] D. Riboni, T. Sztyler, G. Civitarese, and H. Stuckenschmidt, "Unsupervised recognition of interleaved activities of daily living through ontological and probabilistic reasoning," in *Proceedings of ACM UbiComp*. ACM, 2016, pp. 1–12.