Multimed Tools Appl
DOI 10.1007/s11042-013-1419-7

An innovative web-based collaborative platform for video annotation

Isaak Kavasidis · Simone Palazzo · Roberto Di Salvo · Daniela Giordano · Concetto Spampinato

Abstract Large scale labeled datasets are of key importance for the development of automatic video analysis tools as they, from one hand, allow multi-class classifiers training and, from the other hand, support the algorithms' evaluation phase. This is widely recognized by the multimedia and computer vision communities, as witnessed by the growing number of available datasets; however, the research still lacks in annotation tools able to meet user needs, since a lot of human concentration is necessary to generate high quality ground truth data. Nevertheless, it is not feasible to collect large video ground truths, covering as much scenarios and object categories as possible, by exploiting only the effort of isolated research groups. In this paper we present a collaborative web-based platform for video ground truth annotation. It features an easy and intuitive user interface that allows plain video annotation and instant sharing/integration of the generated ground truths, in order to not only alleviate a large part of the effort and time needed, but also to increase the quality of the generated annotations. The tool has been on-line in the last four months and, at the current date, we have collected about 70,000 annotations. A comparative

I. Kavasidis (✉) · S. Palazzo · R. Di Salvo · D. Giordano · C. Spampinato
Department of Electrical, Electronics and Computer Engineering,
University of Catania, Catania, Italy
e-mail: kavasidis@dieei.unict.it

S. Palazzo
e-mail: simone.palazzo@dieei.unict.it

R. Di Salvo
e-mail: roberto.disalvo@dieei.unict.it

D. Giordano
e-mail: dgiordan@dieei.unict.it

C. Spampinato
e-mail: cspampin@dieei.unict.it

Published online: 26 March 2013

performance evaluation has also shown that our system outperforms existing state of the art methods in terms of annotation time, annotation quality and system's usability.

Keywords Ground truth data · Video labeling · Object detection · Object tracking · Image segmentation

1 Introduction

In the last decades, the advancements in camera technology and the reduction of costs of memory storage have led to a proliferation of visual data content in the form of images and videos, and as consequence, a widespread increase in the number of applications for automatic video analysis, such as video surveillance for security, sport and real-life monitoring. For all these purposes, the scientific community has put a lot of effort in the development of powerful object detection [4, 13], tracking [8, 40] and recognition [30] approaches, which, however, are not able to scale up to many scenarios and objects due mainly to the lack of large labeled training sets. Indeed, as demonstrated in other research fields [5, 24], the performance of classifiers increase dramatically when a conspicuous set of labeled training data is available. Moreover, ground truth data plays a central role not only in learning tasks but also in quantitative performance evaluation, which has also received significant attention by the vision community with the aim to establish a valid reference for a systematic evaluation of any computer vision technique.

Therefore, large scale annotated datasets, covering as much scenarios and objects as possible, are needed in order to train and evaluate the existing approaches. The main limitation to achieve this goal is the daunting amount of time and human concentration needed to generate high quality ground truth data, in fact it has been estimated that labeling an image may take from to two to thirty minutes, depending on the operation, and it is, obviously, even worse in the case of videos. In fact, ground truths on videos typically consist of a list of the objects with information on their bounding box, the contour, the recognition class and associations to other appearances in past or following frames.

There exist, in the literature, a few attempts, such as Caltech 101 and 256 [14, 18] and the Berkeley Segmentation dataset [27], produced by some vision groups that have collected consistent annotated datasets, which, however, are too task-oriented and can not be generalized. To reach the objective of creating more diverse and larger annotated datasets, collaborative methods, exploiting large population of expert and motivated users, have been proposed [35, 41]. Beside web-based collaborative solutions, crowdsourcing the annotation efforts to non-experts has been adopted [1, 32]. However, crowdsourcing approaches lack mainly in mechanisms for assessing the reliability of annotators and for combining multiple users' annotations.

Moreover, most of the above approaches are not user-centric, lacking in usability since they are tailored to a specific task but do not adapt to users' needs [38]. In addition, these tools have primarily dealt with the collection of labeled data for human-centered applications (video-surveillance, sports, human behavior understanding), while little has been done on more complex scenarios such as real-life animal monitoring [39], which poses several challenges from a computer vision point

of view, in terms of low (e.g. object detection and tracking) and high processing (e.g. behaviour understanding) levels and the task of video labeling is even harder due to the intrinsic features of the analysed scenes and objects (e.g. in real-life underwater monitoring, scenes are very crowded with many object occlusions and fish move erratically in 3D changing often in appearance and size).

In this paper we propose a web-based approach for video annotation which has two main objectives: 1) to guide and speed up the annotation phase meeting users' needs and 2) to build up a large scale database of labeled visual data to be used in object detection, tracking and image segmentation tasks.

The main contributions of the tool to the research on ground truth collection in multimedia (and in computer vision) are:

- It provides an easy-to-use interface, specific for generating ground truths for object detection, segmentation, tracking and classification.
- It improves the user experience with respect to the existing annotation tools, by showing two panels, each containing a frame at a different time, thus allowing the user to compare a frame's annotations with those from a previous or following frame, and providing quick methods to specify object associations and perform attribute propagation (e.g. hotkeys).
- It integrates effective methods for quality control of the collected data and for the integration of multiple users' annotations.

This tool is, currently, being used to collect large scale ground truth on the real-life underwater video footage of the Fish4Knowledge project[1] which aims at developing automatic video and image analysis methods to support marine biology research.

The remainder of the paper is organized as follows: Section 2 analyzes strengths and limitations of the existing video and image annotation approaches. Section 3, instead, describes the proposed framework, highlighting functionalities and improvements with respect to the state of the art, while, in Section 4, the collected content on the aforementioned underwater environment, is presented. Section 5 shows the performance of the proposed system in terms of 1) the accuracy of the generated ground truths, 2) the efficiency of the platform in ground truth generation and 3) its learnability and user satisfaction. Concluding remarks and future developments are given in Section 6.

2 Related works

Because ground truth generation is a fundamental task in the design and testing of computer vision algorithms, in the last decade the multimedia and, more in general, the computer vision community has developed a disparate number of annotation frameworks and tools to help researchers in collecting datasets, which are then used in the tasks of image segmentation, object detection and tracking, face recognition, etc.

The most common approaches are devised as stand-alone tools created by isolated research groups, and as such, tailored to specific needs. These include, for instance, ViPER-GT [12], GTVT [2], GTTool [23], ODViS [21], which, however, show their

[1] www.fish4knowledge.eu

limitations when it comes to generate large scale ground truth datasets. In fact, they exploit the efforts of a limited number of people and do not support sharing of labeled data. All these needs combined with the rapid growth of the Internet have favored in the last years the expansion of web-based collaborative tools, which take advantage of the efforts of large groups of people to collect reliable ground truths. LabelMe [35], a web-based platform to collect user annotations in still images, is a significant example. However, LabelMe lacks intelligent mechanisms for quality control and integration of user annotations. In fact, quality control is achieved by a simple approach that counts the number of annotation landmarks, and it does not exploit the full potential of its collaborative nature (being a web-based platform) since annotations of multiple users of the same object instance are not combined. In fact, the LabelME dataset, though being one of the largest datasets available, it is particularly inaccurate. Moreover, LabelMe is thought specifically for still images, although a video based version has been proposed [45] that, however, is not as successful and flexible as the image based version.

Sorokin and Forsyth [37] have, recently, demonstrated the utility of "crowdsourcing" to human resources (non-experts) the task of collecting large annotated datasets. Nevertheless, two main aspects have to be taken into account when crowdsourcing: workers' motivation and control. The easiest and most natural way to motivate people is paying them for their work. This strategy is applied by Amazon's Mechanical Turk service [32] and CrowdFlower [7]. A valid alternative for workers motivation is personal amusement [43]: this is the case of the ESP and Peekaboom games [1] which exploit players' agreement (randomly pairing two players and let them guess each other's labels) to collect ground truth data.

Beside workers motivation, another concern of crowdsourcing solutions is the quality control over annotators, which has been tackled with different strategies that can be summarized [31] as: Task Redundancy (ask multiple users to annotate the same data), User Reputation and Groundtruth seeding (i.e. coupling ground truth with test data). Although these solutions are able to build large scale datasets, they might be very expensive and contain low quality annotation since workers (even if payed) are not as motivated as researchers.

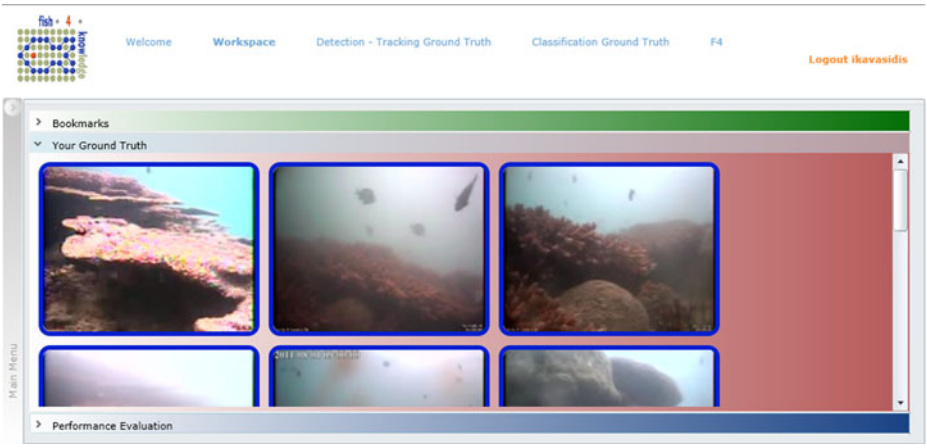

Fig. 1 User workspace

There exist also approaches, which try to generate ground truth data automatically (without human intervention) video and image data [6, 19] but they rely on approaches that cannot be fully trusted.

For all the above reasons, semi-automatic approaches [3, 29, 42], i.e. the ones that involve humans and algorithms in the labeling process, seem to be the most suitable to achieve the goal of large scale ground truth data collection. A few multi-layer class semi-automatic annotation tools have been conceived [15, 36], but they are mainly concerned with static image labeling and, as such, they do not allow analysing how objects behave over time (sequence of frames).

3 The web annotation tool

3.1 General description

The proposed tool[2] is a web-based collaborative environment which allows users to share their own annotations with others accelerating high quality video ground truth generation process by increasing/integrating the number of annotations in a sort of inherent user supervision.

Given an input video stream, our platform extracts video frames and provides a set of utilities to annotate each video frame and to follow objects across frames.

It is a rich internet application, based on a standard client-server architecture: the client is implemented in Silverlight while the server's logic and the communication with the database is developed in C#.

In the next subsections a more detailed description of the proposed application's GUI is given.

3.2 The workspace

Immediately after login, the user is presented with a private workspace where it is possible to review past activities (Fig. 1). The workspace serves also as a shortcut to the labeling operations already performed and is divided in the following sections:

- *Bookmarks*: The use of bookmarks is necessary to reference videos in case of large video collections as the one described in this paper, which contains about half million videos.
- *Ground Truths*: In this section (Fig. 2), the user can manage the ground truths that she owns. In particular, by using the context menu's option a user can create a new ground truth, modify a previously generated one, or derive a new ground truth from an existing one. Moreover, the context menu deals with the collaborative aspect of the platform by allowing the users to make available their ground truths to the other users of the platform. In this way, the users can benefit from the existing annotations and apply the appropriate changes instead of having to generate a new ground truth from scratch.

[2] http://f4k.ing.unict.it/perla.dev

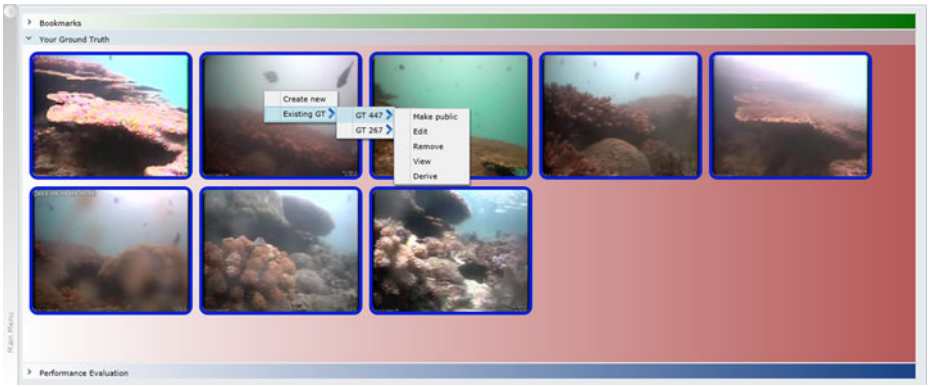

Fig. 2 The ground truth management part of the application. In this section, the videos for which the user created at least one ground truth are included

3.3 Video selection

By clicking on the "Detection—Tracking Ground Truth", the user is presented with the video selection screen (Fig. 3) where it is possible to browse all the available videos, filter them according to specific criteria, bookmark them and start the annotation application. The search engine allows users to limit the number of the shown videos by defining criteria regarding the videos' resolution, acquisition time, enabling the user to select videos with specific features (e.g. day or night) and the exact date of the acquisition.

3.4 Ground truth generation

Once the user identifies the videos she wants to create ground truth for, she can initiate the labeling process by launching the annotation application. This part of the platform permits to create annotations by using multiple windows. Each drawing window (Fig. 4, top left) shows one image and, by using the available toolbox (Fig. 4, bottom), annotations can be drawn on it.

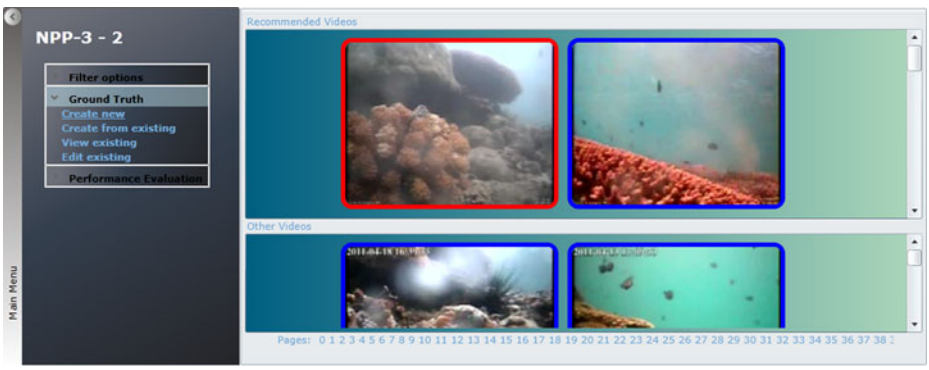

Fig. 3 The video selection window

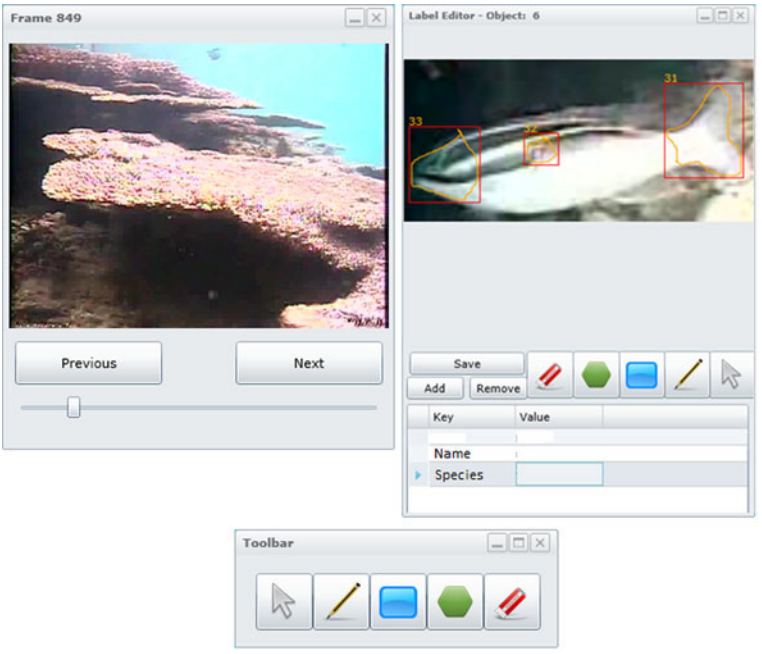

Fig. 4 *Top left*: A drawing window that shows an image to be annotated. The Next and Previous buttons and the slider located at the bottom of the window allow the user to navigate through the video. *Top right*: A labeling window aiming at supporting the user to annotate sub-parts and for adding textual metadata to the designed objects and. *Bottom*: The toolbar. From left to right, the bounding box selection, pencil, rectangle, polygon and eraser tools

3.4.1 Object detection ground truth and contour drawing

The proposed application offers the basic tools (polygon and pencil) to support users in the task of manually drawing object contours. However, manual annotation is discouraging in lengthy image sequences where the numbers are overwhelming. For example, one of the most populated videos in our repository, contained about 18,000 fishes on a 10 min, low resolution, 5 fps videoclip. Under these conditions any means assisting users in drawing object contours as efficiently as possible seems necessary. To this end, the proposed tool implements three automatic contour extraction methods, Grabcut [33], Snakes [22] and Canny [10]. These algorithms were chosen because not only they are well established and tested methods for contour extraction, but also they offer the best ratios in terms of resources and quality of the results. The automatic contour extraction can be applied by drawing the bounding box containing the whole interesting object, right clicking on it and selecting from the "Contour Extraction" sub menu one of the available methods (Fig. 5). This is a trial-and-error process that does not always yield the desired result, because the success of the automatic image contour extraction algorithms depends on the type of image used on (image color patterns, contrast etc.).

In case of automatic contour extraction failure, the user can resort to the manual drawing tools.

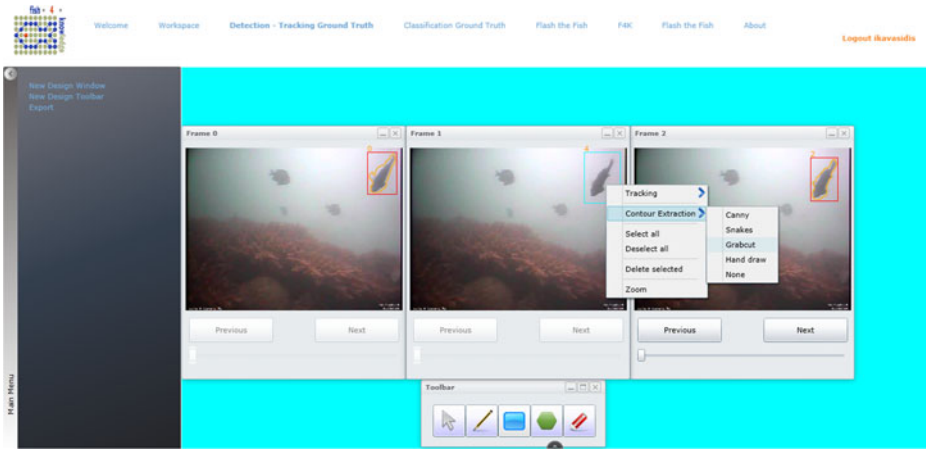

Fig. 5 Semi-automatic contour extraction applied on the center drawing window's image

After an object is drawn, the user can further annotate subparts of it (Fig. 4, Top left) by employing the same tools described above. Furthermore, from the same window, the user is able to add textual metadata to the object (that are included in the exported XML file) that can be useful in other image processing contexts (object recognition, image segmentation, information retrieval etc.).

3.4.2 Object tracking ground truth

In the proposed tool, the tracking ground truth generation exploits the capabilities of multiple windows applications in order to implement an easy-to-use and intuitive way to follow objects across consecutive frames. In particular, to be able to annotate multiple instances of the same object in consecutive frames, the user must arrange side-by-side multiple drawing windows. When the user places two windows with their boarders in direct contact, they become, what we call, a "drawing chain". While chained, the Next and Previous buttons and the sliders of all the drawing windows are disabled except from the last one's (the rightmost), which serves as a control to navigate through the image sequence. Moreover, all the chained windows maintain all the drawing functionalities as if they were unchained. When an adequate, for the user's needs, chain is formed the user must draw an object and bring up the context menu by right clicking on it, then select the voice "Tracking" and select an object from the previous frames she wants to assign the clicked object to (Fig. 6).

When used in high resolution desktop setups, the application can create multiple long chains (as shown in Fig. 7) of successive frames in the same annotation instance (about 3 chains of 6 windows using a 1920×1080 resolution, more on a multi-monitor setup).

3.5 Combining users annotations

The collaborative nature of the proposed tool implies that there may exist multiple annotations of the same object. Such multiple annotations are combined in order to

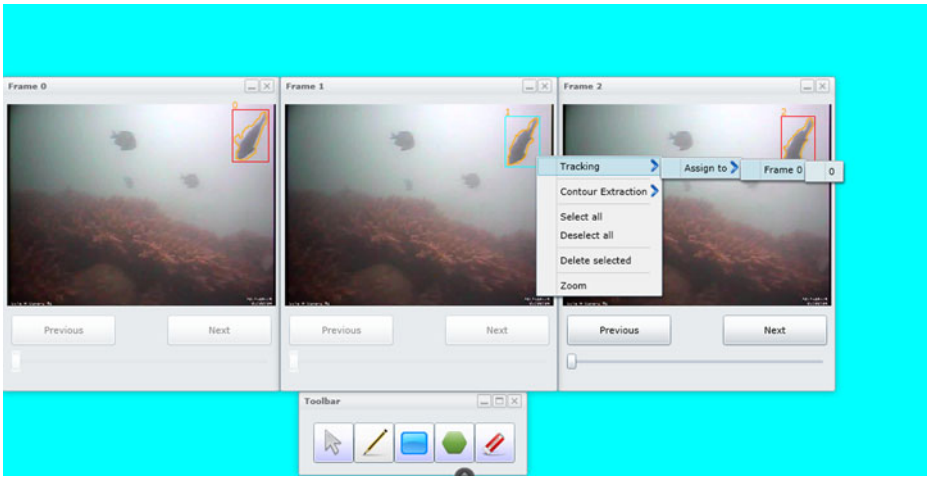

Fig. 6 A three-window chain for tracking ground truth generation

produce a much more accurate object representation since we can safely assume that combined opinions are more objective than single ones [9, 20].

The underlying idea is that for each videoclip we can have more ground truths, annotated by the same or different users, which are integrated by adopting a voting policy in order to generate the one herein called "best ground truth".

The "best ground truth" (BGT) building process (see Fig. 8) involves two basic steps: i) add new annotated objects to the BGT, ii) integrating objects' contours.

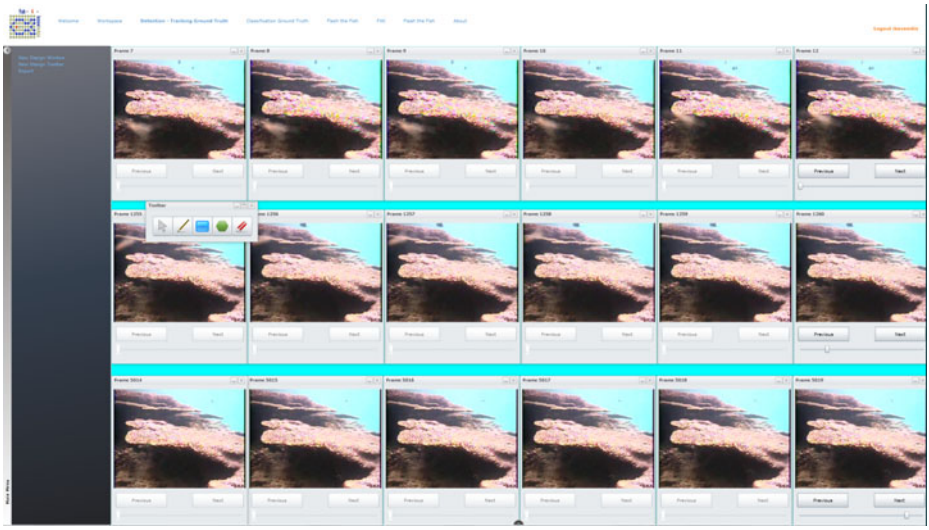

Fig. 7 Three independent six-window chains

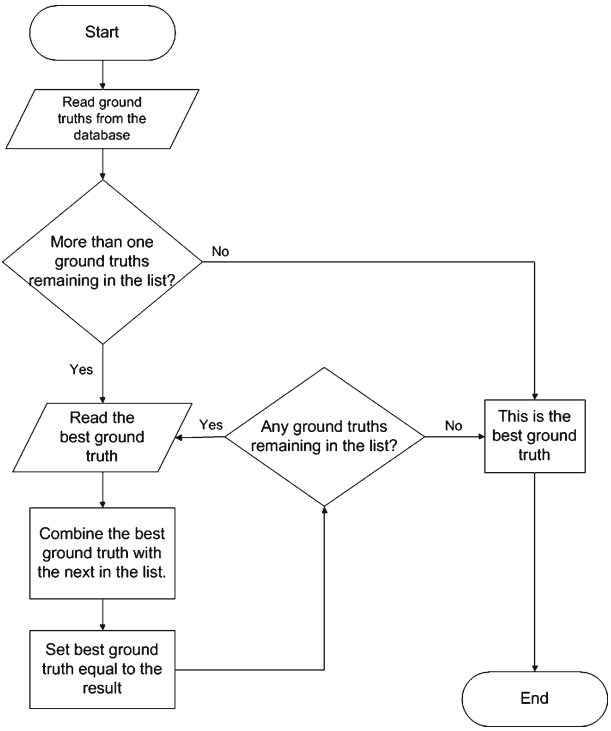

Fig. 8 Flowchart of the "best ground truth" building process

Let us suppose that the BGT has been already built for a given video V and a user annotates V again. For each new annotated object A, two cases may occur:

- **New object instance**. The object A has been never annotated and it is added directly to the BGT. This exploratory strategy avoids limiting the number of objects on each ground truth; however, to prevent noisy ground truths, each object instance in the BGT comes with a number describing the number of annotators that have labeled it over the total number of annotators, thus allowing us to filter out the object instances which have received few annotations.
- **Existing object instance**, i.e. there is already an instance (referred in the following as GT) of object A in the BGT. In this case we assess a matching score between object A and object GT and if this score is greater than a given threshold (in our case 0.75) the contours of A will be combined with the ones of GT. The matching score is computed as weighted mean of the two following measures:
  - Overlap Score. Given the object A and the corresponding object GT of the best ground truth BGT, the overlap score, O_{score}, is given by:

$$O_{\text{score}} = \frac{area(A \cap GT)}{area(A \cup GT)} \tag{1}$$

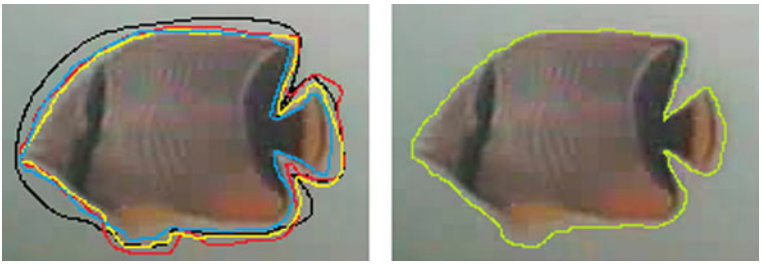

Fig. 9 Combination of annotations for building a "best ground truth" object. On the left there are four annotations: three (black, yellow, red) from different users and one (blue) belonging to the already existing best ground truth. On the right, the resulting contour to be assigned as the new best ground truth

– Euclidean Distance Score. Pairwise euclidean distance between A points (X, Y), with $(X_i, Y_i) \in A$, and GT points (x,y), with $(x_{i'}, y_{i'}) \in GT$, computed as:

$$E_{\text{score}} = 1 - \frac{\sum_i^n \sqrt{(X_i - x_{i'})^2 + (Y_i - y_{i'})^2}}{\max\left(\sum_i^n \sqrt{(X_i - x_{i'})^2 + (Y_i - y_{i'})^2}\right)} \tag{2}$$

Usually, a resampling procedure is applied, in order to equal the number of points in the two contours.

The objects' contours combination is based on the assumption that the "best ground truth" contours are more accurate then the new ones since they result from the combination of multiple annotators. In detail, once a new object is considered for being part of the "best ground truth" (see above) its contours C_A are combined with the contours C_{GT} of the corresponding "best ground truth" object to form the new object contours C_{NGT}, where each point is computed as:

$$C_{\text{NGT}}(i, j) = \frac{1}{2^{N-1}} \sum_{n=1}^{N} (w_A \times C_A(i, j) + C_{GT}(i, j)) \tag{3}$$

where $w_A \in [T, 1]$ (where T is the threshold described above, and is set to 0.75) is the matching score between A and GT computed as above described and N is the number of different annotations for that given object. Figure 9 shows the result of a combination of four annotations (one belongs to the already existing best ground truth) on the same object, whereas Fig. 10 shows how object contours evolve as the number of annotators increases.

Finally, a quality score is assigned to the user (U_{qs}) that represents her ability in ground truth creation, equal to:

$$U_{qs} = \frac{1}{N} \sum_{i}^{N_{GT}} q_i n_i \tag{4}$$

where N is the total number of objects that the user has drawn, N_{GT} is the number of the created ground truths, q_i is the quality of the i_{th} ground truth and n_i is the number of objects belonging to that ground truth.

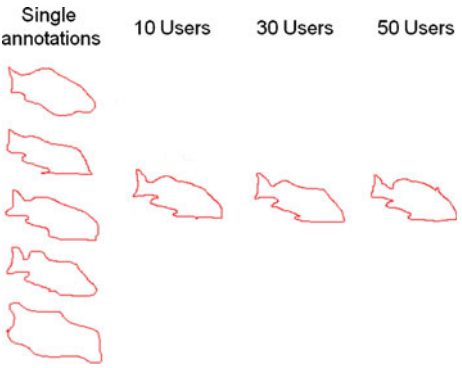

Fig. 10 Object contours quality improvements as the number of annotators gets bigger

4 Data content

The proposed tool has been conceived for ground truth data collection within the Fish4Knowledge project, whose video repository holds more than half a million videos at different resolutions and frame rates. Those videos are acquired by eight high definition underwater cameras that are set to work 24/7.

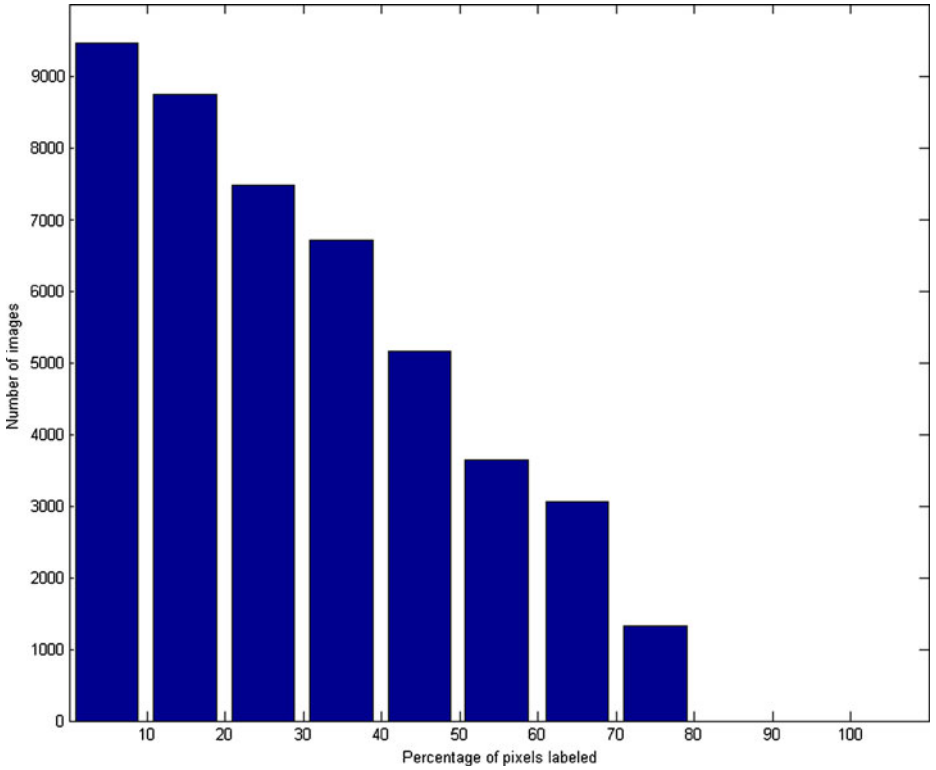

Fig. 11 Histogram of the number of images with respect to the pixel coverage

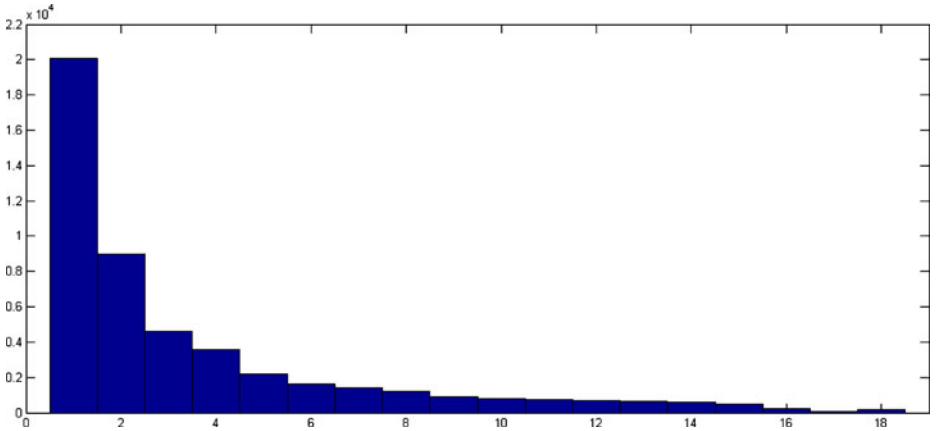

Fig. 12 Histogram of the number of images with respect to the number of objects present

At the date of December 31, our database contains 55 annotated videos with 55,332 annotations (about 2,900 different objects) in 24,136 video frames, collected by several users with our web-based tool, which is online since July 01, 2012.

Figure 11 shows the histogram of the total number of annotated images with respect to the percentage of labeled pixels. In particular, 10,034 frames have less than 10 % of pixels labeled and no image has more than 60 % of pixels labeled. The histogram of the number of images per the number of objects in these images (see Fig. 12), instead, shows that there exists a high number of images with only one annotation (a little more than 11,000).

Currently, the tool's database is constantly growing up, since more and more new users are working on the annotation of new image sequences. At the current rate, we estimate that about 350 10-min videos annotated by the end of 2013, resulting in about 500.000 annotations of about 25.000 different objects.

5 Performance evaluation

The first evaluation of system's performance was carried out in terms of time needed to perform annotations and accuracy of collected annotations. In particular, we asked 50 computer science undergraduate students to annotate fish in 200 consecutive frames of 5 different videos (320 $\times$ 240, 5 *fps*, 10 min long), provided with high quality ground truths, taken from the application's repository, with the proposed tool, the GTTool [23] and ViPER-GT. The users were given a time period of two weeks to complete the task. The time spent on the proposed tool was measured automatically. For the GTTool and ViPER-GT the students were asked to accurately take note of the time spent during the whole process. The achieved results in terms of efficiency and accuracy are shown in Table 1.

The accuracy of the contours was compared against the gold standard ground-truths available for those five videos by calculating the average of the PASCAL score and the Euclidean distance.

The results shown that the time required to annotate the videos on average, was lower for the GTTool. The reason behind this is that the GTTool employs automatic object detection (in addition to automatic contour extraction methods, which are the

Table 1 Comparison between the proposed tool, the GTTool and ViPER-GT

	Proposed tool	GTTool	ViPER-GT
Total drawn objects	34131	43124	31409
Manually drawn objects	16832	14563	31409
Automatically drawn objects	17299	28561	–
Average time per object	7,4 s	4,2 s	11,2 s
Contour accuracy	89 % (95 %)	90 %	79 %
Learnability	9.1	8.2	3.4
Satisfaction	7.3	7.3	4.3

The number in parenthesis is the accuracy obtained by using the contour integration module (see Fig. 13)

same as the proposed tool) to support users' annotations, thus resulting in a major number of automatically drawn objects, as shown in Table 1. For the same reason the accuracy of the annotations drawn with GTTool was also slightly better that the one achieved with the proposed tool. ViPER-GT ranked last in this comparison because of its complete manual nature. It is important, though, to notice that these results refer to a one-vs-one user setting and do not include possible advantages that can be exploited by the proposed tool's multi-user nature.

So, in order to compare the effort needed to generate high quality ground truth by using the aforementioned tools, the time needed to annotate a video containing a hundred fish objects, was measured. In single user applications, such as ViPER-GT and GTTool, which do not offer any annotation integration method, the time necessary to create a ground truth, increases exponentially with respect to its quality. Considering though, that the proposed tool is devised to permit multiple users to collaborate, integrating their annotations gives a significant boost to the quality/effort ratio. In fact, in Fig. 13, is shown the time needed in order to achieve different quality scores. In the single-user case, as it was aforementioned, the best performer is the GTTool, needing about 61 min, in the best case, in order to get a ground truth quality of 0.8. When the annotation integration module was used, the same quality was achieved in about 30 min (in the 50 users setting).

Upon annotation completion, a usability questionnaire [11] was compiled by the users, in order to obtain some feedback about the user experience. In particular, the

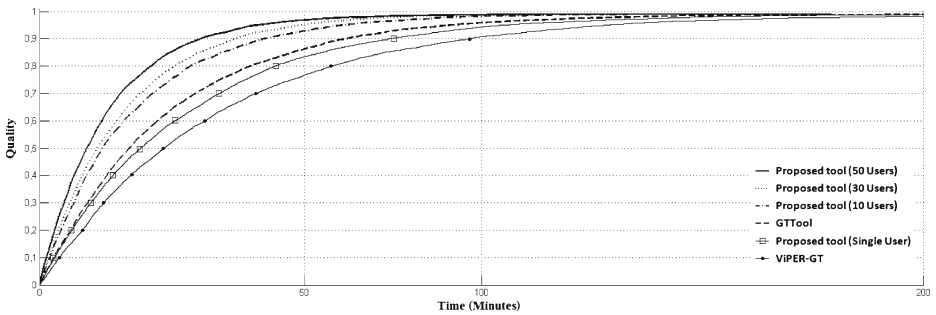

Fig. 13 The time (in minutes) needed and the obtained quality, for annotating a video containing 100 objects. For the single user cases the graphic represents the time needed by the best performing user. For the proposed tool, when annotation integration takes place, it represents the time needed for the users to achieve the corresponding quality score, working in parallel (i.e. the time that each user spent)

students had to grade the used tools in terms both of learnability and satisfaction. Learnability expresses how easy it is to learn to use the tools, while satisfaction represents the general feelings of the users about their time spent with each tool; both values range from 1(worst) to 10 (best).

As shown in Table 1, the totality of users voted the proposed tool as the easiest to learn, achieving a score of 9.1/10, with the GTTool coming near second (8.2/10). ViPER-GT ranked third with a very low score (3.4/10) mainly because of its complex interface and the time needed to achieve a satisfactory level of knowledge on its usage.

When user satisfaction is concerned, both the GTTool and the proposed application achieved a score of 7.3 out of 10. This tie was due to two main reasons, based on the users' comments: 1) The GTTool's object detection and tracking algorithms, alleviated a large part of the work, while 2) the proposed web tool was easier to use, better organized and more appealing to the eye. The worst performer was, again, ViPER-GT because of the total lack of automated tools and its steep learning curve.

6 Concluding remarks

In this paper, a web based video annotation tool is presented. It deals with all the aspects of the ground truth generation process at different vision levels.

Although, the tool is on-line since last July and only few users have had access to it (for testing reasons), about 55.000 annotations have been drawn and we expect this number to grow exponentially in the next months. Besides, the experimental results have shown that the tool presented herein, allows users to speed-up the generation of high quality ground truth due to the distribution of the workload to multiple users. Moreover, the contour integration module performed as expected increasing the quality of the produced ground truth.

Currently we are working on integrating a performance evaluation module which will enable scientists to test and compare their algorithms using the generated ground truth. This module will be further extended with social interaction capabilities, in order to enable scientists to share code, datasets and ideas.

As future developments we plan to add automatic video analysis tools for object detection, tracking, recognition image segmentation that may save annotation time. We also aim to map the currently available XML format into a web ontology, in order to give users the possibility to insert semantic metadata for each annotated object, which could not only support interoperability with other semantic web applications (e.g. multimedia retrieval, like in [17, 25, 34]), but also enable users to generate ground truth for higher level tasks (e.g. object recognition etc.).

Machine learning methods [26, 28, 44] will be applied on these textual annotations in order to exploit the advantages offered by integrating annotations to multiple types and levels of information. These semantic data will be available to the end users via SPARQL Endpoints. The integration of more extensive collaborative capabilities, e.g. simultaneously editing the same ground-truth or systemically distributing the ground truth generation among different users, would undoubtedly accelerate even more the whole process. Moreover, multiple annotations of the same object by different users could be integrated by using Adaboost [16] in order to enhance the quality of the produced ground truth.

Acknowledgements We would like to thank the anonymous reviewers for their constructive and invaluable comments. This research was funded by European Commission FP7 grant 257024, in the Fish4Knowledge project.[3]

References

1. Ahn LV (2006) Games with a purpose. Computer 39(6):92–94
2. Ambardekar A, Nicolescu M, Dascalu S (2009) Ground truth verification tool (GTVT) for video surveillance systems. In: Proceedings of the 2009 second international conferences on advances in computer-human interactions, ser. ACHI '09, pp 354–359
3. Barbour B, Ricanek Jr K (2012) An interactive tool for extremely dense landmarking of faces. In: Proceedings of the 1st international workshop on visual interfaces for ground truth collection in computer vision applications, ser. VIGTA '12. ACM, New York, pp 13:1–13:5
4. Barnich O, Van Droogenbroeck M (2011) ViBe: a universal background subtraction algorithm for video sequences. IEEE Trans Image Process 20(6):1709–1724 [Online]. Available: http://www.ncbi.nlm.nih.gov/pubmed/21189241
5. Bassel GW, Glaab E, Marquez J, Holdsworth MJ, Bacardit J (2011) Functional network construction in Arabidopsis using rule-based machine learning on large-scale data sets. Plant Cell 23(9):3101–3116
6. Bertini M, Del Bimbo A, Torniai C (2005) Automatic video annotation using ontologies extended with visual information. In: Proceedings of the 13th annual ACM international conference on multimedia, ser. MULTIMEDIA '05, pp 395–398
7. Biewald L (2012) Massive multiplayer human computation for fun, money, and survival. In: Proceedings of the 11th international conference on current trends in web engineering, ser. ICWE'11, pp 171–176
8. Blake A, Isard M (1996) The condensation algorithm—conditional density propagation and applications to visual tracking. In: Advances in neural information processing systems. MIT Press, pp 655–668
9. Brabham D (2008) Crowdsourcing as a model for problem solving an introduction and cases. Convergence 14(1):75–90
10. Canny J (1986) A computational approach to edge detection. IEEE Trans Pattern Anal Mach Intell 8(6):679–698
11. Chin JP, Diehl VA, Norman KL (1988) Development of an instrument measuring user satisfaction of the human-computer interface. In: Proceedings of the SIGCHI conference on human factors in computing systems, ser. CHI '88. ACM, New York, pp 213–218
12. Doerman D, Mihalcik D (2000) Tools and techniques for video performance evaluation. In: Proceedings of 15th international conference on pattern recognition, vol 4, pp 167–170
13. Faro A, Giordano D, Spampinato C (2011) Adaptive background modeling integrated with luminosity sensors and occlusion processing for reliable vehicle detection. IEEE Trans Intell Transp Syst 12:1398–1412
14. Fei-Fei L, Fergus R, Perona P (2007) Learning generative visual models from few training examples: an incremental bayesian approach tested on 101 object categories. Comput Vis Image Underst 106(1):59–70
15. Fisher R (2004) CAVIAR test case scenarios. Online Book
16. Freund Y, Schapire RE (1995) A decision-theoretic generalization of on-line learning and an application to boosting. In: Computational learning theory. Springer, pp 23–37
17. Giordano D, Kavasidis I, Pino C, Spampinato C (2011) A semantic-based and adaptive architecture for automatic multimedia retrieval composition. In: 2011 9th international workshop on content-based multimedia indexing (CBMI), pp 181–186
18. Griffin G, Holub A, Perona P (2007) Caltech-256 object category dataset. California Institute of Technology, Tech. Rep. 7694

[3] www.fish4knowledge.eu

19. Heroux P, Barbu E, Adam S, Trupin E (2007) Automatic ground-truth generation for document image analysis and understanding. In: Proceedings of the ninth international conference on document analysis and recognition, ser. ICDAR '07, vol 01, pp 476–480
20. Howe J (2006) The rise of crowdsourcing. Wired Magazine 14(6):1–4
21. Jaynes C, Webb S, Steele R, Xiong Q (2002) An open development environment for evaluation of video surveillance systems. In: PETS02, pp 32–39
22. Kass M, Witkin A, Terzopoulos D (1988) Snakes: active contour models. Int J Comput Vis 1: 321–331
23. Kavasidis I, Palazzo S, Di Salvo R, Giordano D, Spampinato C (2012) A semi-automatic tool for detection and tracking ground truth generation in videos. In: VIGTA '12: proceedings of the 1st international workshop on visual interfaces for ground truth collection in computer vision applications. ACM, New York, pp 1–5
24. Kawahara T, Nanjo H, Shinozaki T, Furui S (2003) Benchmark test for speech recognition using the corpus. In: Proceedings of ISCA & IEEE workshop on spontaneous speech processing and recognition, pp 135–138
25. Mai HT, Kim MH (2013) Utilizing similarity relationships among existing data for high accuracy processing of content-based image retrieval. Multimed Tools Appl. doi:10.1007/s11042-013-1360-9
26. Marques O, Barman N (2003) Semi-automatic semantic annotation of images using machine learning techniques. The Semantic Web-ISWC 2003, pp 550–565
27. Martin D, Fowlkes C, Tal D, Malik J (2001) A database of human segmented natural images and its application to evaluating segmentation algorithms and measuring ecological statistics. In: Proc. 8th int'l conf. computer vision, vol 2, pp 416–423
28. Mihalcea R, Corley C, Strapparava C (2006) Corpus-based and knowledge-based measures of text semantic similarity. In: Proceedings of the national conference on artificial intelligence, vol 21, no 1. AAAI Press, Menlo Park, MIT Press, Cambridge, p 775, 1999
29. Moehrmann J, Heidemann G (2012) Efficient annotation of image data sets for computer vision applications. In: Proceedings of the 1st international workshop on visual interfaces for ground truth collection in computer vision applications, ser. VIGTA '12, pp 2:1–2:6
30. Mutch J, Lowe D (2008) Object class recognition and localization using sparse features with limited receptive fields. Int J Comput Vis 80:45–57
31. Quinn AJ, Bederson BB (2011) Human computation: a survey and taxonomy of a growing field. In: Proceedings of the 2011 annual conference on human factors in computing systems, ser. CHI '11, pp 1403–1412
32. Rashtchian C, Young P, Hodosh M, Hockenmaier J (2010) Collecting image annotations using Amazon's Mechanical Turk. In: Proceedings of the NAACL HLT 2010 workshop on creating speech and language data with Amazon's Mechanical Turk, ser. CSLDAMT '10, pp 139–147
33. Rother C, Kolmogorov V, Blake A (2004) "GrabCut": interactive foreground extraction using iterated graph cuts. ACM Trans Graph 23(3):309–314
34. Rotter P (2013) Relevance feedback based on n-tuplewise comparison and the ELECTRE methodology and an application in content-based image retrieval. Multimed Tools Appl. doi:10.1007/s11042-013-1384-1
35. Russell BC, Torralba A, Murphy KP, Freeman WT (2008) Labelme: a database and web-based tool for image annotation. Int J Comput Vis 77(1–3):157–173
36. Sigal L, Balan A, Black M (2010) Humaneva: synchronized video and motion capture dataset and baseline algorithm for evaluation of articulated human motion. Int J Comput Vis 87(1):4–27. doi:10.1007/s11263-009-0273-6
37. Sorokin A, Forsyth D (2008) Utility data annotation with Amazon Mechanical Turk. In: 2008 IEEE computer society conference on computer vision and pattern recognition workshops. IEEE, Piscataway, pp 1–8
38. Spampinato C, Boom B, He J (eds) (2012) VIGTA '12: proceedings of the 1st international workshop on visual interfaces for ground truth collection in computer vision applications. ACM, New York
39. Spampinato C, Palazzo S, Boom B, van Ossenbruggen J, Kavasidis I, Di R, Salvo Lin F, Giordano D, Hardman L, Fisher R (2012) Understanding fish behavior during typhoon events in real-life underwater environments. Multimed Tools Appl. doi:10.1007/s11042-012-1101-5
40. Spampinato C, Palazzo S, Giordano D, Kavasidis I, Lin F, Lin Y (2012) Covariance based fish tracking in real-life underwater environment. In: International conference on computer vision theory and applications, VISAPP 2012, pp 409–414

41. Stork DG (1999) Character and document research in the open mind initiative. In: Proceedings of the fifth international conference on document analysis and recognition, ser. ICDAR '99
42. Utasi A, Benedek C (2012) A multi-view annotation tool for people detection evaluation. In: Proceedings of the 1st international workshop on visual interfaces for ground truth collection in computer vision applications, ser. VIGTA '12, pp 3:1–3:6
43. von Ahn L, Dabbish L (2004) Labeling images with a computer game. In: Proceedings of the SIGCHI conference on human factors in computing systems, ser. CHI '04. ACM, New York, pp 319–326 [Online]. Available: http://doi.acm.org/10.1145/985692.985733
44. Wache H, Voegele T, Visser U, Stuckenschmidt H, Schuster G, Neumann H, Hübner S (2001) Ontology-based integration of information-a survey of existing approaches. In: IJCAI-01 workshop: ontologies and information sharing, vol 2001. Citeseer, pp 108–117
45. Yuen J, Russell BC, Liu C, Torralba A (2009) Labelme video: building a video database with human annotations. In: ICCV'09, pp 1451–1458

Isaak Kavasidis received the Laurea degree in computer engineering in 2009 from the University of Catania, where he is currently doing his PhD. His interest activities include semantic web, RDF and semantic image processing.

Simone Palazzo received the Laurea degree in computer engineering in 2010, grade 110/110 cum laude from the University of Catania, where he is currently doing his PhD. His interest activities include image and signal processing, image enhancement and reconstruction.

Roberto Di Salvo received the Laurea degree in computer engineering in 2008, grade 110/110 cum laude from the University of Catania, where he is currently doing his PhD. His interest activities include mainly object detection and tracking in real-life scenarios.

Daniela Giordano received the Laurea degree in electronic engineering, grade 110/110 cum laude, from the University of Catania, Italy, in 1990. She also holds the PhD degree in educational technology from Concordia University, Montreal (1998). She is associate professor of information systems at the Engineering Faculty of the University of Catania since 2001. Daniela Giordano's researches have been published in international refereed journals and conference proceedings and has dealt with topics such as: advanced learning technology, knowledge discovery, image processing, and information technology in medicine. Her current research interests include: cognitive systems, multimodal interaction and semantic technologies. She is a member of the IEEE and of the ACM.

Concetto Spampinato received the Laurea degree in computer engineering in 2004, grade 110/110 cum laude, and the PhD in 2008 from the University of Catania, where he is currently research assistant. His research interests include image and signal processing for motion detection systems in environmental applications, image enhancement and reconstruction in bioimaging and content based multimedia retrieval. He has co-authored about 60 publications in international refereed journals and conference proceedings and he is member of The International Association for Pattern Recognition (IAPR).