Implicit Vs. Explicit Human Feedback for Interactive Video Object Segmentation

Francesca Murabito, Simone Palazzo, Concetto Spampinato$^{(\boxtimes)}$, and Daniela Giordano

Pattern Recognition and Computer Vision (PeRCeiVe Lab),
Department of Electric Electronic and Computer Engineering,
University of Catania, Catania, Italy
{fmurabito,palazzosim,cspampin,dgiordan}@dieei.unict.it

Abstract. This paper investigates how to exploit human feedback for interactive object segmentation in videos. In particular, we present an interactive video object segmentation approach where humans can contribute by either explicitly clicking on objects of interest in videos or implicitly while looking at video sequences. User feedback is then translated into a set of spatio-temporal constraints for an energy-based minimization problem. We tested the method on standard benchmarking datasets when using both eye-gaze data and user clicks. The results indicated how our method outperformed existing automated and interactive methods regardless of the type of human feedback (explicit or implicit), and that click-based feedback was more reliable than eye-gaze one.

1 Introduction

The recent progress in digital imaging and smartphone technologies, followed by their relatively low cost and the explosion of social networks, have favoured the generation and sharing of an impressive amount of visual data content over the internet (to give an idea, 80% of all consumer Internet traffic in 2019 will be due to video data[1]). Millions of videos and images are shared daily on Youtube, Facebook, Twitter, Flickr, etc., and now represent the primary source of information and communication.

Nevertheless, this massive visual data can be seen as an added value only if it is possible to analyze and effectively understand it, thus turning raw data into meaningful information needed for several applications: from security to surveillance to ecology monitoring to marketing strategies. This highlights the importance of automated analysis methods that, as a consequence, are proliferating. Unfortunately, such methods are not always capable of satisfying application requirements, especially in terms of expected accuracy. A key role in the understanding process may be played by humans, who, on one hand, have an extraordinary ability in performing high-level tasks with unreachable performance for

[1] http://www.cisco.com/c/en/us/solutions/collateral/service-provider/ip-ngn-ip-next-generation-network/white_paper_c11-481360.html.

S. Battiato et al. (Eds.): ICIAP 2017 International Workshops, LNCS 10590, pp. 131–142, 2017.
https://doi.org/10.1007/978-3-319-70742-6_12

machines, and, on the other hand, have limited processing/computation capabilities, making it impossible to analyze visual data at a large scale. Combining and integrating effectively humans and machines is, therefore, highly desirable, as also witnessed by the recent research front that aims at involving actively humans in the machine learning loop [4,16,20–22,25,29,31].

Through this paper we intend to contribute to the research on humans in the loop for video understanding by attempting to answer the question "what is the most suitable and effective way to exploit human capabilities in analyzing visual data while keeping efforts as low as possible?". This question has a two-fold implication related to human exceptional performance in understanding the visual world: (a) visual scene understanding in humans involves implicit and involuntary processing, such as eye movements, that, if exploited by automated methods, despite being rather noisy information, would allow to reduce significantly human intervention (given the involuntary nature of the feedback); (b) explicitly annotating visual data, e.g., by providing per-frame bounding boxes, is an easy task for humans but extremely tedious and time-consuming, and requires, at large scale, a collective effort.

To support the answer to this research question, we propose an interactive method for video object segmentation, which extends existing interactive methods [2,6,17,26] to work with several interaction modalities converting user feedback into spatio-temporal constraints for the segmentation process. However, the main contribution is the comparison between (a) implicit eye gaze data recorded through an eye-tracker while subjects look at video sequences, and (b) explicit user clicks collected while people play a web game for video object segmentation. We tested our approach on standard video benchmarks and compared the performance of the two interaction modalities, beside comparing it to state-of-the-art automated and interactive video object segmentation methods.

2 Related Work

In this paper we propose a video object segmentation approach posed as a binary labeling task (i.e., background/foreground segmentation) solved through MRF energy minimization as in [8,13,19]. The difference between those methods and ours is that we involve humans in the segmentation loop; therefore, our approach falls within the interactive video segmentation research area [2,6,17]. Interactive video object segmentation methods aim at converting human input (often in the form of drawn lines or strokes) into constraints for spatio-temporal segmentation, so that manual annotations can be propagated to multiple frames. Our paper draws inspiration from these methods but extends them in the way humans and their feedback are included into the video object segmentation process. Indeed, in our work, user feedback is obtained either explicitly by asking multiple users to click on video sequences through a *web-game* or by simply asking single users to look at video sequences and then recording *eye-gaze* data through an eye-tracker.

Games have been already employed for collecting human annotation with the purpose to train and test machine learning methods as interactive segmentation or annotation of images [5,15,23,29,30]. Analogously, eye-gaze has been adopted (a) for identifying the most viewed image regions and their visual descriptors to be used for image tagging [31] and image indexing/retrieval [3] and (b) as a tool for human implicit feedback in a video object segmentation scenario with promising results [27]. As an alternative to eye-gaze, brain activity data recorded through EEG has been utilized as implicit feedback for supporting interactive image annotation [16].

Interactive image and video annotation is an active research area both in multimedia and computer vision, and the existing approaches can be classified into two main categories: (a) methods requiring explicit user feedback as either lines and strokes or user clicks [2,5,6,15,17,23,30], and (b) approaches exploiting implicit user feedback (eye gaze or EEG) [3,16,27,31]. However, the main limitation of these methods is on how user feedback is incorporated into the interactive annotation process, i.e. how to effectively translate user data into spatio-temporal constraints for visual data analysis. Furthermore, most of these methods are thought for image annotation/tagging/segmentation and only few for video object segmentation and, so far, no one of them has compared the performance of implicit vs. explicit feedback.

3 The Interactive Video Segmentation Method

Our interactive video segmentation method is based on [26] with the difference that we make it generic (thus removing some terms which were very application-specific) and able to work with different interaction modalities including eye-gaze data. The whole approach relies on (1) a spatial frame segmentation method which takes into account both visual cues and user input and (b) a spatio-temporal module able to consider spatio-temporal links between image parts in consecutive frames.

3.1 Spatial Frame Segmentation

The first step of the algorithm performs image segmentation at the frame level, which is treated as a binary pixel labeling task (background and foreground). We start from *superpixel* segmentation [1] and then group superpixels through minimization of an energy cost which enforces spatial and visual coherency between superpixels as well as including user feedback, as constraints, in the labeling cost. The underlying idea is that superpixels "selected" by users (either by clicking on them or by simply looking at them) should be defined as hard constraints in the energy minimization problem.

Let $P = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be user feedback in the form of (x, y) points for frame F. We define an energy function over the set of F's superpixels S able to model superpixels "selected" by users, and at the same time, to enforce spatial constraints on visual smoothness at the frame level. The energy function for spatial segmentation is based on the assumptions that superpixels identified by multiple users can be considered as hard constraints for segmentation as well as unselected superpixels that are spatially-close and visually-similar to selected ones; and single superpixels should be ignored as possibly noisy. In particular, it is defined as:

$$E_1(\mathcal{L}) = \alpha_1 \sum_{s \in S} F_1(s, l_s, P) + \sum_{(s_1,s_2) \in \mathcal{N}(S)} F_2(s_1, s_2, l_{s_1}, l_{s_2}) \tag{1}$$

where $\mathcal{L} = \left\{ l_{s_1}, l_{s_2}, \dots, l_{s_{n_S}} \right\}$ is the superclick label assignment (l_{s_i} is the binary superclick label for superpixel s_i), $\mathcal{N}(S)$ is the set of pairs of neighbor superpixels (that is, having part of boundary in common; we will also use the notation $\mathcal{N}(s)$ to denote the set of neighbors of the single superpixel s), and α_1 is a weighing factor.

F_1 takes into account if a superpixel s should be part of an object or not. As this depends on how many users have selected it and on neighboring superpixels, it is given by two contributions:

- ***User feedback*** f_s **on superpixel** s**:** the more a superpixel has been selected by users, the more it is likely to be an object part. The score f_s for superpixel s is:

$$f_s = \frac{|P \cap s|}{\max\limits_{t \in S} |P \cap t|} \tag{2}$$

  where $P \cap s$ is the set of user data hitting superpixel s and $|\cdot|$ is set cardinality. The term takes into account how many times superpixel s has been selected by users.
- **Adjacency** A_s**:** if superpixel s has not been selected by users but it is adjacent to superpixels which did, it is safe to consider it as foreground as well.

  The proximity term A_s is computed as the fraction of neighbor superpixels with $I_{s_n} > \theta$, with $s_n \in \mathcal{N}(s)$ and $\theta = 0.6$:

$$A_s = \frac{|\{s_n \in \mathcal{N}(s) : I_{s_n} > \theta\}|}{|\mathcal{N}(s)|} \tag{3}$$

The sum of f_s and A_s (clipped to 1 if necessary) is the likelihood that a superpixel s is part of the foreground objects, $P_{s,1} = \min(f_s + A_s, 1)$, while $P_{s,0} = 1 - P_{s,1}$ is the probability that s is "not a part" of the foreground. In the energy function E_1, F_1 encodes the cost of assigning a certain label to each superpixel and is the negative log-likelihood of $P_{s,1}$ and $P_{s,0}$:

$$F_1(s, l_s, C) = \begin{cases} -\log P_{s,1} & \text{if } l_s = 1 \\ -\log P_{s,0} & \text{if } l_s = 0 \end{cases} \tag{4}$$

F_2 is instead the cost of assigning different labels to two adjacent and visually similar superpixels s_1 and s_2 and in our case is computed as:

$$F_2(s_1, s_2, l_{s_1}, l_{s_2}) = KL(H_{s_1}, H_{s_2}) \mathcal{I}(l_{s_1} \neq l_{s_2}) \tag{5}$$

where KL is Kullback-Liebler distance, H_{s_i} is the RGB color histogram of superpixel s_i, and $\mathcal{I}$ is an indicator function which returns 1 if the arguments is true, and 0 otherwise. The per-frame segmentation is then obtained by minimizing $E_1(\mathcal{L})$ through graph cut; examples are given in Fig. 1 (first row).

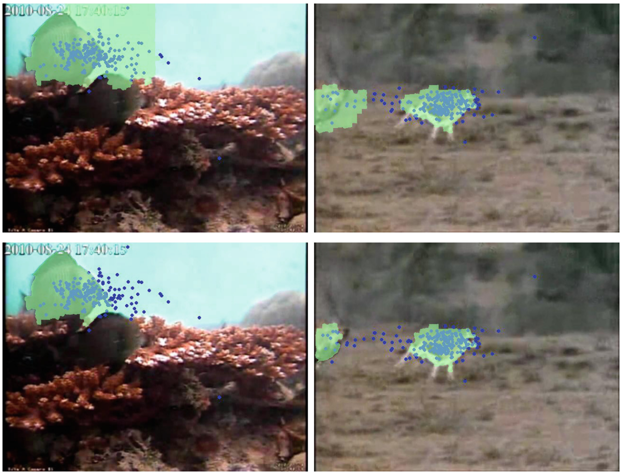

Fig. 1. Output examples: (first row) segmentation masks when using only spatial information; (second row) segmentation refinement by including temporal constraints. Blue dots are user input in the form of clicks in the example. (Color figure online)

3.2 Spatio-Temporal Segmentation

The previous step converts user-provided (x, y) points into a set of potential foreground superpixels, but it does not take into account any temporal link between superpixels in consecutive frames, which, instead, is necessary to cope with per-frame segmentation errors. To do that we used the idea proposed in [26] which is based on the assumption that if a set of adjacent superpixels is selected in consecutive frames, then it is very likely that it is part of an object. Nevertheless, superpixel extraction is often not consistent in presence of large motion. This aspect is addressed by including a temporal-based segmentation part, which links superpixels in consecutive frames according to their visual similarity; and makes an hypothesis on the position of superpixels in consecutive frames through optical flow [14]. More specifically, superpixels are linked in consecutive frames by introducing pairwise potentials on all pairs of superpixels $\{s_t, s_{t+\xi}\}$ such that s_t contains at least one pixel p_t whose projection $p_{t\to t+\xi}^{v_{p_t}} = p_t + v_{p_t}$ into frame $t+\xi$ under the motion vector v_{p_t} (i.e., v_{p_t} is the motion vector computed between frame t and frame $t+\xi$ for location p_t) is part of superpixel $s_{t+\xi}$ in frame $t+\xi$. Thus, we did not consider only linking between two consecutive frames as in [26] since user feedback can be faster than clicks as in the case of eyes.

The energy term encoding spatio-temporal constraints uses (as in the original work) a batch of $2T+1$ consecutive frames from $t-T$ to $t+T$ and is defined as:

$$\begin{aligned} E_2(\mathcal{L}) = & \sum_{\tau=t-T}^{t+T} \left[\sum_{s \in S^\tau} F_1(s, l_s, l_s^\tau) \right] + \\ & + \sum_{\tau=t-T}^{t+T} \left[\sum_{(s_1,s_2) \in \mathcal{N}(S^\tau)} F_2(s_1, s_2, l_{s_1}, l_{s_2}) \right] + \\ & + \sum_{(s_1,s_2) \in \mathcal{N}_T(\cup_{\tau=t-T}^{t+T} S^\tau)} F_2(s_1, s_2, l_{s_1}, l_{s_2}) \end{aligned} \tag{6}$$

The first two lines of Eq. 6 are, respectively, a unary potential for each identified superpixel (first line) and a pairwise potential for each pair of superpixels belonging to the same frame (second line). The last term (third line), instead, aims at enforcing temporal smoothing through a pairwise potential defined over the set $\mathcal{N}_T(\cup_{\tau=t-T}^{t+T} S^\tau)$, i.e. the set of all pairs of superpixels in the $2T+1$ "temporally linked" frames, as described above. F_1 models whether superpixel s is more likely to be background or foreground: if s was labeled as foreground in the frame-segmentation we expect it to be more likely that it is foreground (with a cost lower than being background), and vice versa. F_1 is therefore computed as:

$$F_1(s, l_s, l_s^\tau) = \begin{cases} \gamma_1 & \text{if } (l_s = 1 \wedge l_s^\tau = 1) \vee (l_s = 0 \wedge l_s^\tau = 0) \\ \gamma_2 & \text{otherwise} \end{cases} \tag{7}$$

with $\gamma_1 < \gamma_2$.

F_2 is computes the similarity between "temporally-adjacent" superpixels in consecutive frames as in Sect. 3.1. E_1, E_2 are minimized through graph cut. Segmentation examples are shown in the second row of Fig. 1: compared to those of the first row, these examples highlight how including temporal-based refinement enhances segmentation performance.

4 Experimental Results

Experiment settings. We tested our method using two user interaction modalities: (1) eye-gaze data recorded by a Tobii T60 eye tracker (with a capture frequency of 60 Hz) while subjects looked at video sequences, and (2) user clicks collected through a web game on the same set of videos.

The eye-gaze experiments involved sixteen (16) subjects, who were asked to watch a set of short videos with the goal of following moving objects. For the click-based experiment we re-adapted (and used the source code released by the authors) the web-based game proposed in [11], changing only the displayed video sequences and leaving the underlying gamification strategy unchanged. In practice, the game consists of several levels, with each level displaying one or more video sequences. Twelve (12) users were asked to click on moving objects and, accordingly, were rewarded with a score reflecting click accuracy.

Datasets and baselines. Performance evaluation and comparison between the two considered interaction modalities was carried out on 9 video sequences, with pixelwise annotations, taken from three challenging visual benchmarks for video object segmentation: SegTrack v2 [12], FBMS-59 [18], and VSB100 [7]. The selected videos included features such as: camera motion, slow object motion, object-background similarity, non-rigid deformations and articulated objects. The comparison between our approach and automated methods was performed over the Youtube-Objects dataset [24]. The automated video object segmentation methods were tested using public source code and default parameters. As for interactive segmentation methods, we did not perform any accuracy comparison since, to the best of our knowledge, all of them require interaction times not compatible with large scale analysis. For example, labeling 10 video frames enough to achieve an F_1 accuracy of 0.7 took about 50 s with our approach, and more than 1,000 s with [17].

Collected data. Each of the 12 subjects for the web game experiment spent approximately 9.5 min playing the game, which resulted in the collection of 40.4 clicks per frame, on average, for a total engagement time of 115 min. Instead, each eye-gaze experiment took 2 min per subject, and provided on average 35.7 points per frame in 32-min user engagement time. This difference in the amount of time required to collect similar amounts of data was expected, due to the higher acquisition rate of the eye-tracker compared to human clicking speed.

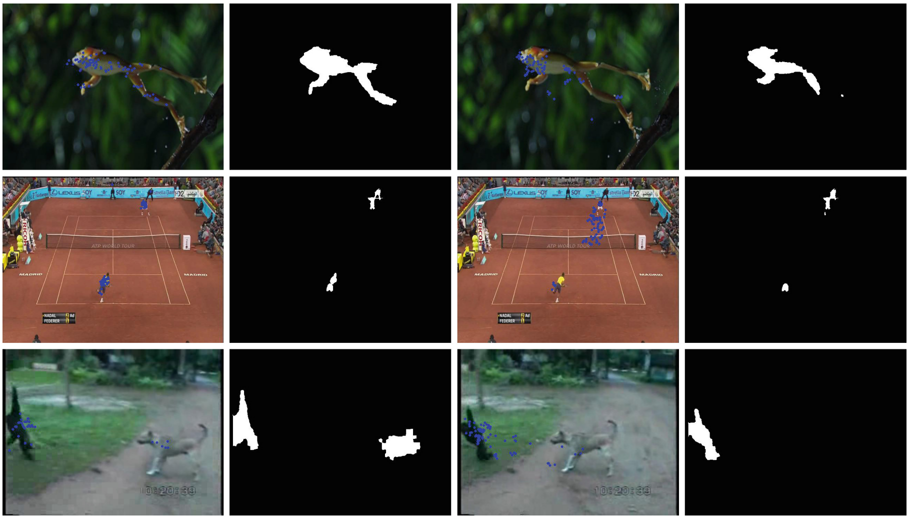

Fig. 2. Segmentation output examples. (First two columns) User clicks and (second two columns) and eye-gaze on sample frames and related output segmentation masks. (Color figure online)

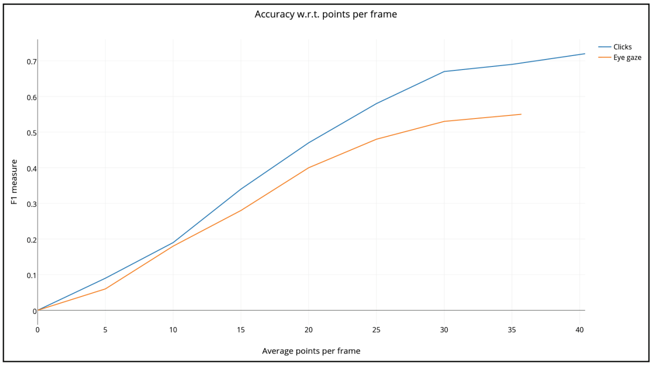

Fig. 3. F1 accuracy w.r.t. user feedback points per frame.

Results and discussion. Table 1 reports pixelwise average F$_1$-measure achieved by our method using different feedback modalities. It is possible to see how click-based feedback (see Clicks$_M$ column in Table 1) outperformed the eye-gaze–based one; a visual comparison in terms of output segmentation masks between click and eye-gaze–based interaction is given in Fig. 2.

The primary reason behind the lower performance of the eye-gaze–based approach lies in the noisy nature of eye movements: eye gaze, indeed, involves both fixations and saccades with the latter spanning the whole visual scene from fixation to fixation. Thus, in case of isolated objects it tended to perform fairly well (fixations and saccades were close) while in case of multiple objects it failed. Higher performance of the click-based approach was also due to the following:

- In both experiments, top-down saliency was enforced since all participants were instructed to follow moving objects; however, in the web-game, user behaviour was driven by game rewards and consequently by competition among players, which was a strong and effective incentive to click on objects accurately;
- Since users were allowed to play the game several times, after the first time they had a prior knowledge on object location. To account for this aspect, we assessed segmentation accuracy using only the clicks generated by the 12 subjects the first time they played the game. The achieved performance is shown in column Clicks$_S$ of Table 1: it is possible to notice how the comparison between the click-based approach (Clicks$_S$) and eye-gaze based one was more balanced, thus indicating that object location prior is a key factor for accurate results.

We also investigated how the number of user feedback points (either clicks or eye fixations) affected the segmentation performance. Figure 3 shows how the F_1 measure changed w.r.t. points per frame. When few points were available, the difference between the two interaction modalities was small, while when the number of points available became consistent their performance diverged significantly. This reinforces our previous claim about the noisy nature of eye-gaze data especially when dealing with multiple objects; moreover, it confirms that when users are driven by a proper incentive to carry out a specific task, performance improves.

Additionally, it can be interesting to see how the proposed method, albeit more in line with the research on interactive video annotation, compares to state-of-the-art automated video object segmentation approaches. To do that, we used the Youtube-Objects dataset (largely employed as a benchmark for video object segmentation), and compared our approach (using game clicks as data points) with a selection of recent methods exploiting superpixels for the video object segmentation, namely, [9,10,18,19,28,32]. The results, in terms of average *Pascal Overlap Measure* (POM, i.e., intersection over union between output masks and ground truth segmentation masks) in percentage, are reported in Table 2, and show that our method outperforms automated video object segmentation methods. In general, this is not surprising, since it is common that interactive video annotation tools perform better than automated methods, but, on the contrary, they are hardly usable in case of large video datasets (e.g., Youtube-objects). While our method can be seen as an interactive video object segmentation approach, it enables multiple users to co-operate in large scale tasks (for the web-game, it might suffice to publish it on a social network and make people play to collect enough data) reducing the annotation/interaction burden, which usually lies on the shoulders of few people. Furthermore, the performance achieved in this work (72.8) was better than the ones in [26] (68.9) with much less users (12 vs. 63 players). This was due to the removal of several terms including the assessment of user quality or the assessment of superpixel similarity, and,

Table 1. F-measure scores obtained by the proposed method, using either eye-gaze data or user-click data. As for user-click we performed two evaluations: (a) exploiting clicks of first-time-play by users in order to eliminate the bias due to prior knowledge on object location (column Clicks$_S$), and (b) using all collected clicks (column Clicks$_M$).

Video	Gaze	Clicks$_S$	Clicks$_M$
animal_chase (VSB)	0.62	0.26	0.78
sled_dog_race (VSB)	0.52	0.53	0.76
tennis (VSB)	0.26	0.57	0.62
cheetah (SegT)	0.67	0.52	0.68
frog (SegT)	0.57	0.57	0.74
monkeydog (SegT)	0.43	0.40	0.64
camel01 (FBMS)	0.65	0.52	0.59
rabbits02 (FBMS)	0.70	0.72	0.89
rabbits04 (FBMS)	0.50	0.76	0.77
Average	0.55	0.54	0.72

Table 2. POM in percentage for the Youtube-Objects dataset

	[18]	[28]	[9]	[32]	[19]	[10]	Ours
Aeroplane	13.7	17.8	73.6	75.8	70.9	86.3	68.4
Bird	12.2	19.8	56.1	60.8	70.6	81.0	64.3
Boat	10.8	22.5	57.8	43.7	42.5	68.6	66.7
Car	23.7	38.3	33.9	71.1	65.2	69.4	72.5
Cat	18.6	23.6	30.5	46.5	52.1	58.9	61.4
Cow	16.3	26.8	41.8	54.6	44.5	68.6	77.2
Dog	18.0	23.7	36.8	55.5	65.3	61.8	76.4
Horse	11.5	14.0	44.3	54.9	53.5	54.0	87.0
Motorbike	10.6	12.5	48.9	42.4	44.2	60.9	80.3
Train	19.6	40.4	39.2	31.4	29.6	66.3	74.1
Average	15.5	23.9	46.3	53.7	53.8	67.6	72.8

provide indications that suitable changes to the segmentation method combined to increased motivation of subjects leads to better performance.

5 Conclusions

In this paper we presented a general interactive video object segmentation approach able to work with different user interaction modalities. We tested it on challenging video sequences by employing either eye gaze or user clicks as human feedback. The conclusions that can be drawn from performance analysis are: (1) task-driven user clicks allow for accurate segmentation performance; (2) collecting user clicks from multiple users is not enough to yield good performance, and prior knowledge on object location proved to be an influencing factor, and (3) eye-gaze user interaction allows for greatly reducing interaction times at the expenses of segmentation accuracy. In the future, we plan to perform a large-scale evaluation involving many more users as well as video sequences for a more accurate analysis of interaction behaviour in order to discovery which visual descriptors are mainly employed by users and incorporate such features into automated methods.

References

1. Achanta, R., Shaji, A., Smith, K., Lucchi, A., Fua, P., Susstrunk, S.: SLIC superpixels. EPFL Technical report 149300, p. 15, June 2010
2. Badrinarayanan, V., Galasso, F., Cipolla, R.: Label propagation in video sequences. In: 2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, pp. 3265–3272 (2010)

3. Buscher, G., Dengel, A., van Elst, L.: Eye movements as implicit relevance feedback. In: CHI 2008 Extended Abstracts on Human Factors in Computing Systems, CHI EA 2008, pp. 2991–2996. ACM, New York (2008). https://doi.org/10.1145/1358628.1358796
4. Deng, J., Krause, J., Fei-Fei, L.: Fine-grained crowdsourcing for fine-grained recognition. In: CVPR 2013, pp. 580–587 (2013)
5. Druck, G., Settles, B., McCallum, A.: Active learning by labeling features. In: EMNLP 2009, pp. 81–90. Association for Computational Linguistics, Stroudsburg (2009). http://dl.acm.org/citation.cfm?id=1699510.1699522
6. Fathi, A., Balcan, M.F., Ren, X., Rehg, J.M.: Combining self training and active learning for video segmentation. In: Proceedings of the British Machine Vision Conference 2011, pp. 78.1–78.11 (2011). http://www.bmva.org/bmvc/2011/proceedings/paper78/index.html
7. Galasso, F., Nagaraja, N., Cardenas, T., Brox, T., Schiele, B.: A unified video segmentation benchmark: annotation, metrics and analysis. In: IEEE International Conference on Computer Vision (ICCV), December 2013. http://lmb.informatik.uni-freiburg.de/Publications/2013/NB13
8. Giordano, D., Murabito, F., Palazzo, S., Spampinato, C.: Superpixel-based video object segmentation using perceptual organization and location prior. In: Computer Vision and Pattern Recognition (2015)
9. Godec, M., Roth, P.M., Bischof, H.: Hough-based tracking of non-rigid objects. In: 2011 International Conference on Computer Vision, pp. 81–88, November 2011
10. Jain, S.D., Grauman, K.: Supervoxel-consistent foreground propagation in video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part IV. LNCS, vol. 8692, pp. 656–671. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10593-2_43
11. Kavasidis, I., Spampinato, C., Giordano, D.: Generation of ground truth for object detection while playing an online game: productive gaming or recreational working? In: 2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, pp. 694–699, June 2013
12. Lee, Y.J., Kim, J., Grauman, K.: Key-segments for video object segmentation. In: ICCV 2011, pp. 1995–2002 (2011). https://doi.org/10.1109/ICCV.2011.6126471
13. Lim, J., Han, B.: Generalized background subtraction using superpixels with label integrated motion estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 173–187. Springer, Cham (2014). https://doi.org/10.1007/978-3-319-10602-1_12
14. Liu, C., Adviser-Freeman, W., Adviser-Adelson, E.: Beyond pixels: exploring new representations and applications for motion analysis. In: Proceedings of the 10th European Conference on Computer Vision, Part III, pp. 28–42 (2009)
15. Maji, S.: Discovering a lexicon of parts and attributes. In: Fusiello, A., Murino, V., Cucchiara, R. (eds.) ECCV 2012. LNCS, vol. 7585, pp. 21–30. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33885-4_3
16. Mohedano, E., Healy, G., McGuinness, K., Giró-i Nieto, X., O'Connor, N.E., Smeaton, A.F.: Object segmentation in images using eeg signals. In: ACM MM 2014, pp. 417–426. ACM, New York (2014). https://doi.org/10.1145/2647868.2654896
17. Nagaraja, N.S., Schmidt, F., Brox, T.: Video segmentation with just a few strokes. In: IEEE International Conference on Computer Vision (ICCV), December 2015
18. Ochs, P., Malik, J., Brox, T.: Segmentation of moving objects by long term video analysis. IEEE PAMI **36**(6), 1187–1200 (2014)

19. Papazoglou, A., Ferrari, V.: Fast object segmentation in unconstrained video. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 1777–1784 (2013)
20. Parikh, D., Crandall, D., Grauman, K.: Discovering localized attributes for fine-grained recognition. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3474–3481, June 2012
21. Parikh, D., Zitnick, C.L.: Human-debugging of machines. In: Neural Information Processing Systems, pp. 1–5 (2011)
22. Parkash, A., Parikh, D.: Attributes for classifier feedback. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012, Part III. LNCS, vol. 7574, pp. 354–368. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33712-3_26
23. Peng, B., Zhang, L., Zhang, D., Yang, J.: Image segmentation by iterated region merging with localized graph cuts. Pattern Recogn. **44**, 2527–2538 (2011). http://www.sciencedirect.com/science/article/pii/S0031320311001282, semi-Supervised Learning for Visual Content Analysis and Understanding
24. Prest, A., Leistner, C., Civera, J., Schmid, C., Ferrari, V.: Learning object class detectors from weakly annotated video. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition, pp. 3282–3289, June 2012
25. Salvador, A., Carlier, A., Giro-i Nieto, X., Marques, O., Charvillat, V.: Crowd-sourced object segmentation with a game. In: Proceedings of the 2nd ACM International Workshop on Crowdsourcing for Multimedia - CrowdMM 2013, pp. 15–20 (2013)
26. Spampinato, C., Palazzo, S., Giordano, D.: Gamifying video object segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **PP**(99), 1 (2016)
27. Spampinato, C., Palazzo, S., Murabito, F., Giordano, D.: Using the eyes to "see" the objects. In: Proceedings of the 23rd ACM International Conference on Multimedia, MM 2015, pp. 1231–1234. ACM, New York (2015). https://doi.org/10.1145/2733373.2806324
28. Tang, K., Sukthankar, R., Yagnik, J., Fei-Fei, L.: Discriminative segment annotation in weakly labeled video. In: Proceedings of International Conference on Computer Vision and Pattern Recognition (CVPR 2013) (2013)
29. Vijayanarasimhan, S., Grauman, K.: Large-scale live active learning: training object detectors with crawled data and crowds. Int. J. Comput. Vis. **108**(1–2), 97–114 (2014)
30. Von Ahn, L., Liu, R., Blum, M.: Peekaboom: a game for locating objects in images, pp. 55–64 (2006)
31. Walber, T., Scherp, A., Staab, S.: Can you see it? Two novel eye-tracking-based measures for assigning tags to image regions. In: Advances in Multimedia Modeling, vol. 7732, pp. 36–46 (2013)
32. Zhang, Y., Chen, X., Li, J., Wang, C., Xia, C.: Semantic object segmentation via detection in weakly labeled video. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR), June 2015