Hierarchy of Localized Random Forests for Video Annotation

Naveen Shankar Nagaraja, Peter Ochs, Kun Liu and Thomas Brox

Computer Vision Group
University of Freiburg, Germany
{nagaraja@informatik.uni-freiburg.de}

Abstract. We address the problem of annotating a video sequence with partial supervision. Given the pixel-wise annotations in the first frame, we aim to propagate these labels ideally throughout the whole video. While some labels can be propagated using optical flow, disocclusion and unreliable flow in some areas require additional cues. To this end, we propose to train localized classifiers on the annotated frame. In contrast to a global classifier, localized classifiers allow to distinguish colors that appear in both the foreground and the background but at very different locations. We design a multi-scale hierarchy of localized random forests, which collectively takes a decision. Cues from optical flow and the classifier are combined in a variational framework. The approach can deal with multiple objects in a video. We present qualitative and quantitative results on the Berkeley Motion Segmentation Dataset.

1 Introduction

An annotated video carries rich information which can be used in many tasks such as improving object classifiers, action recognition and pose estimation. Annotating a video manually is a time consuming task and it comes at a cost. Therefore, researchers have been looking at ways to automate parts of this process [20]. Unsupervised large scale video segmentation in general scenes is currently beyond our abilities. However, making a manual segmentation available in one frame determines the objects we are interested in and tells about their appearance. Automatic segmentation of the rest of the video then becomes a tractable problem, as shown in Fig. 1.

The main challenge in this task is the variation of the considered objects, particularly in combination with disocclusion phenomena. As new parts of an object become visible, there is no direct counterpart in the annotated frame. Consequently, we must learn an object representation from the single annotated frame that generalizes over the typical variations.

For this purpose, we propose the use of a hierarchy of localized random forests trained on the initial frame. In contrast to a global classifier, a localized classifier allows to distinguish objects that have similar global statistics, but differ locally. For instance, some of the glove's dark texture in Fig. 3(a) is similar to parts of the background, but *locally* the textures are never the same. As the optimum scale of localization is not known *a priori*, we suggest combining multiple scales

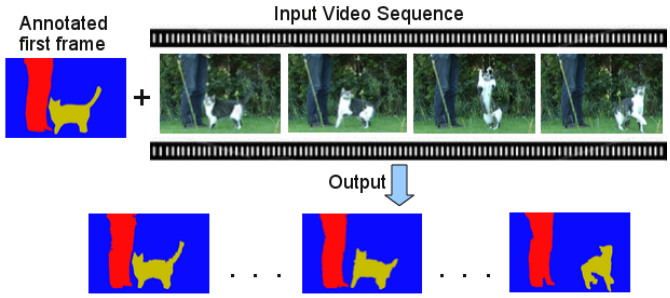


Fig. 1. Given an annotated first frame of a video sequence, we want to achieve a dense labeling of the entire video by propagating the annotations (labels) in the first frame.

in a hierarchy of classifiers. Random forests are particularly well suited for this purpose, as they are very efficient and naturally allow combining results from multiple forests.

Pixel classification is only needed in areas that cannot be directly propagated via optical flow. We combine these two complementary cues in a variational framework, where an additional regularizer ensures compact solutions. We show quantitative results on the Berkeley Motion Segmentation Dataset [5], which provides pixel accurate ground truth.

2 Related work

There are many problem settings of video segmentation in the literature. Interactive video segmentation [14, 12] relies on frequent user intervention to prevent the errors from being propagated. It is very accurate and appreciated in video editing, yet the interactive user input prevents its application to large scale video annotation.

Our work is also related to classical tracking methods. Usually a bounding box is propagated to later frames. Boosting based classifiers [17] and random forest classifiers [16] are popular choices for learning and updating the template model. Godec et al. [10] also give a rough segmentation of the object being tracked. Also level set tracking [9] yields rough segmentations, yet with a less sophisticated appearance model. Moreover, these methods are restricted to tracking a single region.

It is also quite common to consider video segmentation as a spatio-temporal MRF optimization problem [2, 18, 7]. Some of the other methods make use of the superpixels. They connect them spatially and temporally to generate temporally consistent object regions [15, 4, 19, 11]. This approach is often used in an unsupervised setting, which usually leads to severe over-segmentation. As superpixels ignore weak object boundaries, large errors can occur.

In [13] a variant of optical flow is used to propagate labels directly from a training image to a test image. This is similar to how we use optical flow to propagate labels from one frame to the next. Applied over large frame distances,

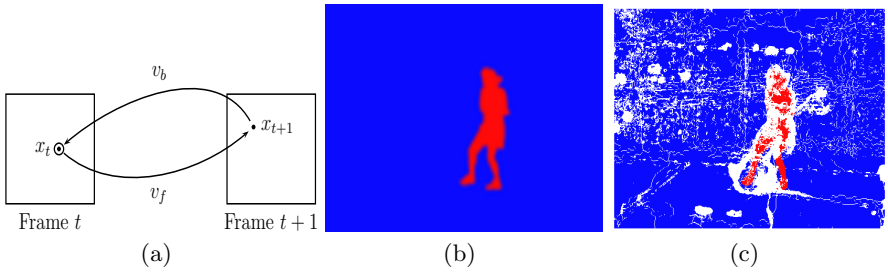


Fig. 2. (a) Forward-backward flow consistency: let a pixel in frame t be denoted by x_t and let the forward flow from frame t to t+1 be v_f. Then, $x_{t+1} = x_t + v_f(x_t)$. Let the backward flow be v_b. If there is no occlusion or estimation error then $x_{t+1} + v_b(x_{t+1}) = x_t$. As the optical flow is never perfectly accurate, we allow for a small radius r, in which we consider the flow as reliable: $\| v_f(x_t) + v_b(x_t + v_f(x_t)) \| \leq r$. **(b)** and **(c)** Labels propagated from frame t to t+1. The white areas denote pixels with unreliable flow.

this leads to erroneous results due to disocclusion. In contrast, we combine the concept of propagation by optical flow with a localized classifier that is trained to generalize over the object's appearance and can fill disocclusion areas.

3 Optical Flow for Frame-wise label propagation

We start with a user segmented first frame of a given video sequence. Let n different class labels be present. The labels are transferred from a frame at time t to t+1 using optical flow from [5]. Optical flow is not reliable near object and motion boundaries, particularly as there may be disocclusions from frame t to t+1. It is important to exclude such critical areas from propagation. They are detected by checking the consistency of the forward and the backward flow, as explained in Fig. 2a.

Let x_t denote a pixel in frame t and its corresponding pixel in frame $t+1$ be x_{t+1}. If the flow passes the consistency check, we assign x_{t+1} the label at x_t. Refer Fig.2(c) for a sample result. Areas that did not pass the consistency check are marked in white. When applying the propagation successively, these areas will get larger and larger, if they are not filled by complementary information.

4 Patch Classification using Localized Classifiers

For each pixel in frame $t+1$ which has not been assigned a label by optical flow, we take the features of a patch around it and feed it to a classifier that has been trained on the annotated first frame. We use random forests [3] since they are computationally efficient, have good generalization properties and can be easily applied to multi-class problems. Moreover, results from multiple random forests can be combined in an elegant manner. This property will prove useful in the following combination of multiple local classifiers. As features we use normalized

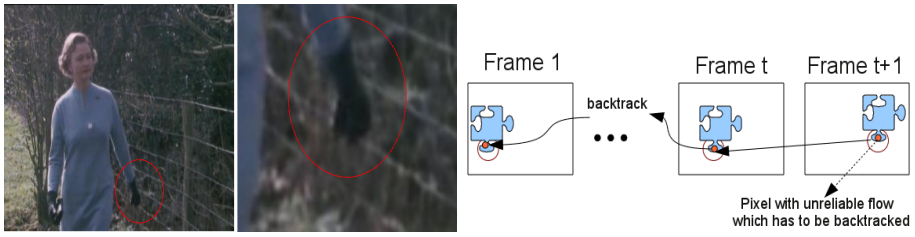


Fig. 3. (a) Advantage of localized classifier over a global classifier- the encircled area can be easily misclassified as background by a global classifier whereas a local classifier for an image block can be more discriminative. **(b)** Coarsely backtracking a pixel with unreliable flow to the annotated Frame 1. A superpixel region around this pixel (shown inside the circle) is backtracked using the average flow within the region.

LAB color histograms computed on 15×15 patches. Each channel is treated separately with 32 bins per channel.

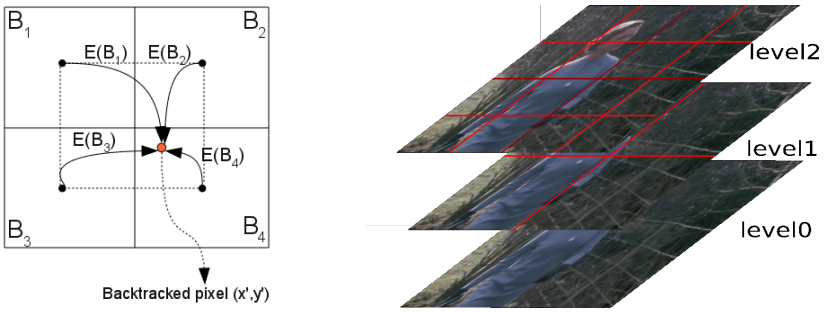
4.1 Localization of the Classifier

A single global classifier comes with the problem that overlapping appearance characteristics between the classes lead to misclassifications. We propose the use of localized classifiers based on the observation that the class overlap is less likely to occur with respect to the local appearance distribution; see Fig. 3a. We exploit the 'locality of reference' by dividing the image into non-overlapping spatial blocks and training a separate random forest in each block. At the test time, we run into a technical problem though: as objects move, the location of an area of interest will be different from that in the training image. Consequently, we must model the shift of the location to choose the right local classifier. We take a region around the pixel that should be classified and recursively backtrack it to the annotated frame; see Fig. 3b. Since we cannot backtrack the pixel itself - the optical flow was unreliable at this point - we use the average flow of the reliable flow vectors in a larger neighborhood around that pixel. To avoid neighborhoods that span different objects, the neighborhood is defined by a superpixel computed with the method in [1] using [8]. The size of the superpixel - steerable by choosing the level in the superpixel hierarchy of [1] - is adapted such that it comprises at least one reliable flow vector.

Even though the backtracking is far from being exact, the resulting location (x', y') is accurate enough to choose the right local random forest. To avoid block artifacts, we run bilinear interpolation on the output of the random forests in the vicinity of (x', y'), as shown in Fig. 4(a).

4.2 Hierarchy of Classifiers

An important question is how to choose an optimal block size. Intuitively it makes sense to have a multi-scale model, where random forests are trained for



(a) (b)

Fig. 4. (a) Interpolation of the classification decisions by the participating blocks at a particular hierarchy level. **(b)** Hierarchy of Localized Random Forests- a Random Forest is trained for each block at each hierarchy level.

different block sizes and their decisions are combined to obtain a final score. Apart from the advantage that we avoid a block size parameter, the additional averaging over multiple forests has positive effects on the overall classification performance.

We use a 3-level hierarchy with one random forest trained on the whole image ($level_0$), 2×2 blocks at $level_1$, and 4×4 blocks at $level_2$, as shown in Fig. 4(b). We use 100, 30 and 10 trees for each forest at $level_0$, $level_1$ and $level_2$, respectively.

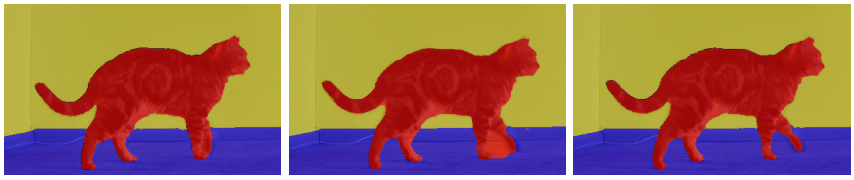
Instead of giving each level of the hierarchy the same influence, we suggest a weighted average with a weight based on the entropy of the respective random forest. Minimum entropy is already used as a splitting criterion when training a random forest. Now we use the entropy in the opposite way: a high entropy in a block indicates a good mixture of labels from all the classes. Thus the classifier trained in a high entropy block will have better data to make a decision than a classifier that has just seen a single label, i.e., a low entropy block. Let $B^l = \{B_1^l, ..., B_4^l\}$ denote the set of neighboring blocks of (x', y') used for bilinear interpolation at hierarchy level l. We have the frequencies of all the labels in a block B_j^l as $p^1(B_j^l), ..., p^n(B_j^l)$, where n denotes the number of class labels. The entropy of a block is then,

$$h(B_j^l) = -\sum_{i=1}^{n} p^i(B_j^l) \log(p^i(B_j^l)) \tag{1}$$

Let $s^i(B_j^l)$ be the score of label i output by the Random Forest for block B_j^l, then the combined score for label i is,

$$s_i = \sum_{j,l} \alpha_j^l exp(h(B_j^l)) s^i(B_j^l) \tag{2}$$

where α_j^l is the weight due to the bilinear interpolation.



(a) (b) (c)

Fig. 5. (a) Frame t. **(b)** shows labels propagated using only optical flow, the "drag" effect is evident from the segmentation near object motion, i.e., the legs. **(c)** shows the effect of our framework in addition to the optical flow to get a more accurate segmentation.

5 Integration of the cues into a Variational framework

In the previous sections we have described two different sources of information to label each pixel independently by either optical flow based label propagation or a classifier decision. Additionally, we must avoid noisy decisions. Hence, we need a smoothness prior that prefers a homogenous labeling in smooth image areas. All of this can be integrated nicely in a variational approach.

Let L_i be the set of coordinates occupied by label i transferred by optical flow and $\mathbf{u}' := (u'_1, \dots, u'_n) \colon \Omega \to \{0,1\}^n$, $n \in \mathbb{N}$ be a function indicating the n different labels, i.e.,

$$u'_i(x) := \begin{cases} 1, & \text{if } x \in L_i \\ 0, & \text{else,} \end{cases} \tag{3}$$

where $\Omega \subset \mathbb{R}^2$ denotes the image domain.

We seek a function $\mathbf{u} := (u_1, \dots, u_n) \colon \Omega \to \{0,1\}^n$ that stays close to the labels propagated from the previous frame using optical flow for points in $L := \bigcup_{i=1}^n L_i$. This is achieved by minimizing the energy

$$E_{\mathrm{OF}}(\mathbf{u}_{t+1}(x)) := \frac{1}{2} \int_\Omega c(x) \sum_{i=1}^n \Big(u_{i,t+1}(x) - u'_{i,t}(x + v_b(x)) \Big)^2 dx, \tag{4}$$

where v_b is the backward flow from frame $t{+}1$ to t and and $c \colon \Omega \to \{0,1\}$ is a confidence indicator function with value 1 where the optical flow passes the consistency check and 0 elsewhere.

On points not covered by optical flow based propagation, we make use of the local classifier's output by introducing one more data term which will aid the decision process at these points to take a specific label. We define this part of our energy functional as

$$E_{\mathrm{LC}}(\mathbf{u}_{t+1}(x)) := \int_\Omega (1 - c(x)) \sum_{i=1}^n -s_i(x) u_{i,t+1}(x) dx, \tag{5}$$

where s_i is the score given by the classifier (2) at position x for label i. The advantage of adding (5) over just combining (4) with the smoothness prior is shown in Fig. 5.

A convex combination of the data terms along with a regularizer yields

$$E := \alpha(E_{\mathrm{OF}} + E_{\mathrm{LC}}) + (1-\alpha)E_{\mathrm{Reg}} \tag{6}$$

s.t. $\sum_i u_{i,t+1}(x) = 1,\ \forall x$ with a model parameter α in $[0, 1)$.

We use a weighted TV based regularizer,

$$E_{\mathrm{Reg}}(\mathbf{u}_{t+1}(x)) = \int_\Omega f(x) \sqrt{\left(\sum_{i=1}^{n} |\nabla u_{i,t+1}(x)|^2 + \varepsilon^2\right)} dx, \tag{7}$$

where $f(x)=(|\nabla I(x)|^2 + \varepsilon^2)^{\frac{-1}{2}}$ is a weighting function that reduces smoothing across image edges.

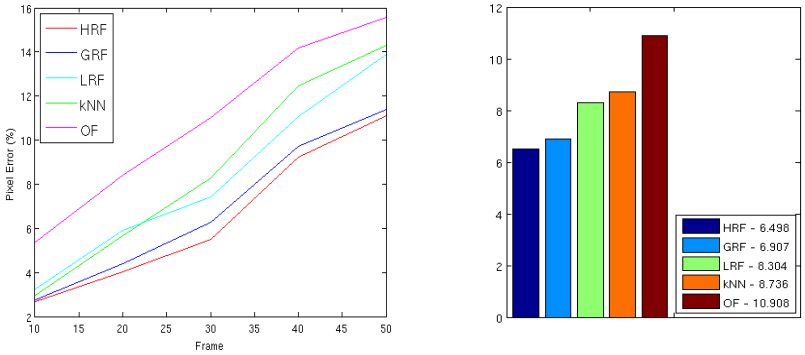
For minimizing (6) we relax the indicator functions u_i to take values in $[0, 1]$. The relaxed problem is convex and we obtain its global optimum by computing the Euler-Lagrange equations and solving the nonlinear system via fixed point iterations and successive over-relaxation (SOR). We obtain the final integer solution by assigning each x the label i with the maximum $u_i(x)$.

6 Experiments and Results

We present quantitative results on the Berkeley Motion Segmentation Dataset (BMS) [6]. BMS has 26 challenging video sequences with varying contrast and illumination and significant motion. For quantitative analysis we have evaluated our output against the ground truth at frame 10, 20, 30, 40 and 50. The evaluation is done by measuring the percentage of pixels that are assigned a different label than the ground truth. This pixel error is calculated for only those classes which were present in the annotated first frame, i.e., new objects that appear later in the video are not taken into consideration. We have kept the parameters fixed for all the sequences with $\alpha = 0.6$. Our method takes approximately 75 seconds to segment each frame on a standard computer. The running time includes the optical flow and superpixel computation as well as the SOR solver (which occupies a major percentage of the running time).

Fig. 6 shows the average pixel error across all sequences in BMS. A comparison to the baseline, which uses only optical flow to propagate the labels and fills the unknown areas with the smoothness prior (i.e., setting $E_{\mathrm{LC}} = 0$), reveals the importance of adding a classifier that deals with disocclusion areas. In particular when labels are propagated over many frames, a classifier improves results considerably.

The comparison between a global random forest and a k-nearest-neighbor classifier further shows that the discriminative training of random forests clearly outperforms a simple memorization of the patches of the first frame. Apart



(a) (b)

Fig. 6. (a) Average pixel error over the BMS sequences after 10,20,30,40 and 50 frames. We compare the proposed hierarchy of random forests (HRF) to a baseline using only optical flow (OF), a k-nearest-neighbor classifier (kNN), a random forest trained on the whole image (GRF) and random forests trained on the lowest level of hierarchy (LRF) i.e. 4×4 blocks. **(b)** The bar chart shows the average pixel error over all the sequences over all the 50 frames with the exact value in the legend.

from the better performance, random forests are also much faster than a k-NN classifier.

Finally, the comparison between the global random forest, the localized random forest on 4×4 blocks, and the proposed hierarchy shows that the hierarchy performs the best. Fig. 7 shows some qualitative results including some typical failure cases.

7 Conclusion

We have presented a method for video segmentation in which we start with a segmented frame and propagate its labels by combining two complementary cues in subsequent frames - one based on optical flow and another based on classification. For the classifier we have proposed a hierarchy of localized random forests, which outperforms the classical global random forests as shown by a quantitative evaluation on a publicly available dataset. The point-wise label predictions are combined with a smoothness prior in a variational setting. The approach can deal with disocclusion and appearance changes as well as multiple objects. It can cover 30-50 frames even on challenging material without introducing large errors. We believe that this tool will help in providing fully segmented video material, which will be useful in many computer vision tasks.

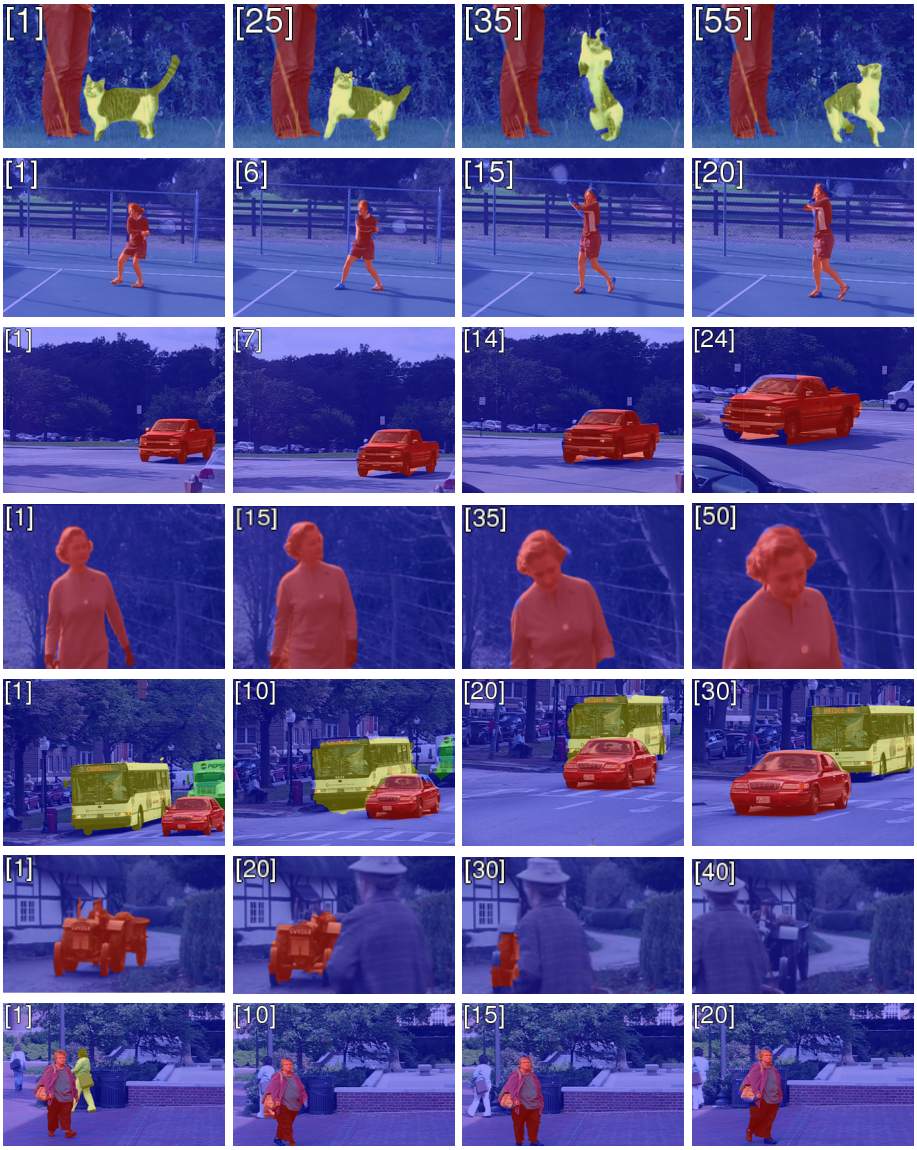


Fig. 7. Segmentation results for some of the BMS sequences with their corresponding frame number. The last two rows show typical failure cases, where an object gets fully occluded by another.

Acknowledgements

We gratefully acknowledge the partial funding by the ERC Starting Grant - VIDEOLEARN.

References

1. Arbelaez, P., Maire, M., Fowlkes, C., Malik, J.: Contour detection and hierarchical image segmentation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2011)
2. Badrinarayanan, V., Galasso, F., Cipolla, R.: Label propagation in video sequences. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
3. Breiman, L.: Random forests. Machine Learning 45, 5–32 (2001)
4. Brendel, W., Todorovic, S.: Video object segmentation by tracking regions. IEEE International Conference on Computer Vision (ICCV) (2009)
5. Brox, T., Malik, J.: Large displacement optical flow: descriptor matching in variational motion estimation. IEEE Transactions on Pattern Analysis and Machine Intelligence (2010)
6. Brox, T., Malik, J.: Object segmentation by long term analysis of point trajectories. European Conference on Computer Vision (ECCV) (2010)
7. Budvytis, I., Badrinarayanan, V., Cipolla, R.: Semi-supervised video segmentation using tree structured graphical models. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2011)
8. Catanzaro, B., Su, B., Sundaram, N., Lee, Y., Murphy, M., Keutzer, K.: Efficient, high-quality image contour detection. International Conference on Computer Vision (ICCV) (2009)
9. Chockalingam, P., Pradeep, N., Birchfield, S.: Adaptive fragments-based tracking of non-rigid objects using level sets. IEEE International Conference on Computer Vision (ICCV) (2009)
10. Godec, M., Roth, P., Bischof, H.: Hough-based tracking of non-rigid objects. IEEE International Conference on Computer Vision (ICCV) (2011)
11. Grundmann, M., Kwatra, V., Han, M., Essa, I.: Efficient hierarchical graph-based video segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2010)
12. Li, Y., Sun, J., Shum, H.Y.: Video object cut and paste. ACM Trans. Graph. (2005)
13. Liu, C., Yuen, J., Torralba, A.: Nonparametric scene parsing- label transfer via dense scene alignment. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2009)
14. Price, B.L., Morse, B.S., Cohen, S.: Livecut: Learning-based interactive video segmentation by evaluation of multiple propagated cues. IEEE International Conference on Computer Vision (ICCV) (2009)
15. Ren, X., Malik, J.: Tracking as repeated figure/ground segmentation. IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2007)
16. Saffari, A., Leistner, C., Santner, J., Godec, M., , Bischof, H.: On-line random forests. ICCV'09 Workshop on On-line Computer Vision (2009)
17. Stalder, S., Grabner, H., Gool, L.V.: Beyond semi-supervised tracking: Tracking should be as simple as detection, but not simpler than recognition. ICCV09 Workshop on On-line Learning for Computer Vision (2009)
18. Tsai, D., Flagg, M., Rehg, J.M.: Motion coherent tracking with multi-label mrf optimization. British Machine Vision Conference (BMVC) (2010)
19. Vazquez-Reina, A., Avidan, S., Pfister, H., Miller, E.: Multiple hypothesis video segmentation from superpixel flows. European Conference on Computer Vision (ECCV) (2010)
20. Yuen, J., Russell, B., Liu, C., Torralba, A.: Labelme video- building a video database with human annotations. IEEE International Conference on Computer Vision (ICCV) (2009)