

Object flow

Anonymous ACCV 2014 submission

Paper ID ***

Abstract. Motion analysis in image sequences has undoubtedly shown good progress in terms of its two main research branches. Optical flow estimation and object visual tracking have been mostly studied as isolated problems, and high accuracy algorithms are available when needed as independent bricks. This paper presents a framework for combining object tracking techniques with optical flow methods aiming towards a precise motion description for objects in video sequences. Firstly, we introduce a method to extend max-flow min-cut based segmentation techniques to videos, without adding the computational load of performing a graph cut optimization approach for 3-dimensional graphs. This is done by exploiting the inherent foreground-background separation hints given by object trackers, and the novel concept of superpixel flow. Then, we show that long-motion awareness obtained from object tracking, together with a per frame object segmentation can improve the precision of the object motion description in comparison to several optical flow techniques. We may call the proposed approach Object flow as it offers a dense and semantic aware description of the current motion state of the studied object.

1 Introduction

Object tracking and optical flow are two of the main components in the computer vision toolbox, and have been focus of great research efforts, leading to significant progress in the last years [16][17]. The object tracking problem consist on estimating the position of the target in future frames, given an initialization. In the other hand, the optical flow between a pair of frames consist on finding a motion vector for each pixel of interest in the initial image. Even though for several applications a full motion-field is needed, other applications like human-computer interaction, object editing in video or structure-from-motion, may only focus on an interest object and, thus, only motion vectors within its space may be of interest. In such scenarios combining optical flow and object tracking in a unified framework would become useful and the precision of the object motion description could be enhanced. For instance, even with modern optical flow approaches, the long term motion problem remains a challenge. However, the problem is more bearable for object tracking techniques. In contrast, object trackers are more global in motion description, and its information can be completed by optical flow sub-pixel precision. Moreover, even when object trackers and optical flow could give good hints for object segmentation in video, these

2 ACCV-14 submission ID ***

elements are not deeply studied in the literature as a unified problem. We introduce the object flow problem as the computation of dense motion flow fields of the set of pixels that belong to an interest object. In other words, the object flow by definition induces the segmentation of the target and its motion field.

We can define more precisely the object flow by starting with an image sequence and an initial position of the interest object in the first frame of this sequence, and letting \mathcal{R} be the region corresponding to the support of the object in 2D, such that $\mathcal{R} \subset \Omega$. If Ω is the set of all the possible grid positions, the object flow problem consist in finding the displacement vector $d_{0,t}(x)$ from the image I_0 to I_t , $\forall x \in \mathcal{R}$.

A straightforward solution to this problem would be to compute the optical flow motion field, and apply a segmentation mask to recover the desired motion vectors. Nevertheless, this approach carries several problems. For example, a globally computed optical flow method can affect small objects motion, because of the common use of heavy regularization priors. Moreover, even if the segmentation mask is extracted from a tracker position by a graph-cut based method, is likely that this mask is not going to be well suited for the interest object and some extra user interaction would be needed to refine this process. We propose an approach to reduce these problems.

The present paper is organized as follows. We describe our pipeline for object flow, including the novel concept of superpixel flow in Sec. 2 and its use in object segmentation in videos. In following sections some results showing how the object flow overpass state of the art optical flow methods for object motion flows estimation are discussed. Finally, some insights and conclusions are given.

2 Algorithm description

The Fig. 1 shows a simplified block diagram of the proposed system. Two details are important, the use of the tracker window to initialize a segmentation procedure, and the use of this segmentation over the tracked window to perform a more precise motion flow computation in the interest pixels. The dotted line represents the possible interaction between precise flow information with the next tracker state. For instance, the current object flow can work as direction hint, and the segmentation information can be used to improve the sampling process of the learning stage in several trackers by detection methods [22], and thus the tracker and motion flow algorithm can work for mutual enhancement.

The first step in the object flow pipeline can be selected according to specific need for a given application. We prefer, in general, tracking-by-detection methods like *Struck* [22] or *MIL* [23], but other approaches could be followed. In the second place, for the object segmentation in video we propose the use of labelled background regions through the concept of superpixel flow, which is explained in the next section. Finally, the flow motion field is computed within the segmentation boundaries and long term dense point trajectories may be extracted from this.

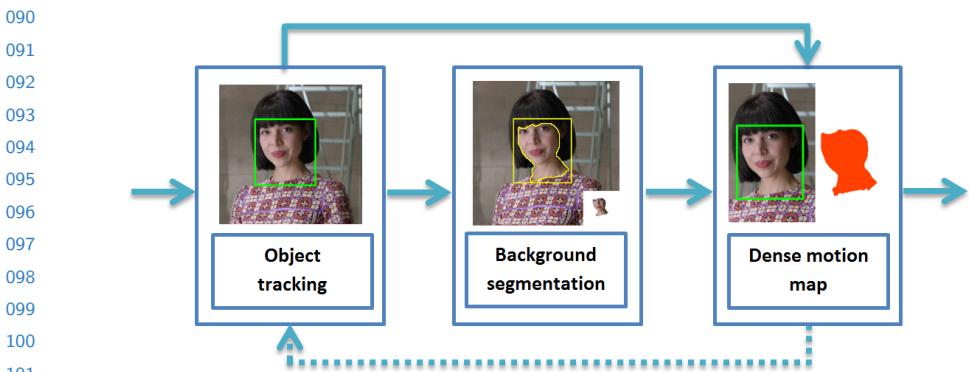


Fig. 1. Block diagram of the proposed pipeline.

2.1 Superpixel flow

As a preprocessing step in the object flow pipeline, we propose a superpixel matching technique which assumes a flowlike behavior in the image sequences (natural video), which can be used to track superpixels. This matching, however, has to comply with a set of constraints. Firstly, two correspondent superpixels should be similar in terms of some appearance feature, which most likely depends on the way the superpixelization was performed (color, texture, shape). Also, the superpixel flow should maintain certain global regularity (at least for superpixels that belong to the same object). If the size compactness of the superpixels is maintained, it actually seems to share some of the properties of the optical flow problem, with the difference that the smoothness is usually a very strong constraint for the last one. The strength of this smoothness prior relies not only in the nature of the problem, but also because it gives better cues towards an easier-to-minimize global approach.

The objective of the superpixel flow is therefore to find the best labeling l for every superpixel p (with $l_p \in 0, 1, \dots, N - 1$) between a pair of frames (I_0, I_1) , but holding a flow-like behaviour.

Thus, the superpixelization should maintain certain size homogeneity within a single frame. Some super pixel techniques can cope with this requirement [9][10]. For the experiments presented in this work, we prefer the SLIC method [9], which usually gives good results in terms of size homogeneity and compactness of the superpixelization.

Inspired by a large number of optical flow and stereo techniques [7][12][13], the superpixel flow can be modelled with pairwise Markov Random Fields. If the matching is performed with MAP inference, its energy function extracted from the posterior probability is:

$$E(l) = \sum_{p \in \Omega} D_p(L_p; I_0, I_1) + \sum_{p, q \in \mathcal{N}} S_{p,q}(L_p, L_q) \quad (1)$$

```

090
091
092
093
094
095
096
097
098
099
100
101
102
103
104
105
106
107
108
109
110
111
112
113
114
115
116
117
118
119
120
121
122
123
124
125
126
127
128
129
130
131
132
133
134

```

135 With l the set of labels of the super pixels in I_0 , that match with those in I_1 .
 136 \mathcal{N} is a neighbourhood of the superpixel p , which defines its adjacency. Given this
 137 posterior probability, the equivalent energy function can be directly obtained by
 138 extracting the negative logarithm of the posterior,

139 The terms D , and S in (1) stand for data term and spatial smoothness terms
 140 as they are popularly known in the MRF literature. The first one determines
 141 how accurate is the labeling in terms of consistency of the measured data (color,
 142 shape,etc.). In the classical optical flow formulation of this equation, the data
 143 term corresponds to the pixel brightness conservation[7][5]. However, as super-
 144 pixels are a set of similar (or somehow homogeneous) pixels, an adequate color
 145 based feature can be a low dimensional color histogram. So D can be written
 146 more precisely as the Hellinger distance between the histograms:

$$147 \quad D_p(l_p; I_0, I_1) = \sqrt{1 - \frac{1}{\sqrt{h(p)h(p')}N^2} \sum_i \sqrt{h_i(p)h_i(p')}} \quad (2)$$

148 Where $h(p)$ and $h(p')$ are the histograms of the superpixel p and its cor-
 149 respondent superpixel in the second frame I_1 . Note that the low dimensional
 150 histogram gives certain robustness against noise, and slowly changing colors be-
 151 tween frames.

152 In the other hand, the spatial term is a penalty function for horizontal and
 153 vertical changes of the vectors that have origin in the centroid of the superpixel
 154 of the first frame and end in the centroid of the superpixel of the second frame.

$$155 \quad S_{p,q}(l_p, l_q) = \lambda(p) \sqrt{\frac{|u_{p_c} - u_{q_c}|}{\|p_c - q_c\|} + \frac{|v_{p_c} - v_{q_c}|}{\|p_c - q_c\|}} \quad (3)$$

$$156 \quad \text{where, } \lambda(p) = (1 + \rho(h(p), h(q)))^2$$

157 In (3) the operator ρ is the Hellinger distance as used in the data term (2).
 158 The histogram distance is nonetheless computed between superpixels p and q ,
 159 which belong to the same neighbourhood. The superpixels centroids are noted
 160 as q_c and p_c , and u and v are the horizontal and vertical changes between cen-
 161 troids. This term is usual in the MRF formulation and has a smoothing effect in
 162 superpixels that belong to the same object. It has to be observed that when two
 163 close superpixels are different, thus, more probable to belong to different objects
 164 within the image, the term λ allows them to have matches that do not hold the
 165 smoothness prior with the same strength. It has to be noted that the proposed
 166 energy function is highly non-convex.

167 The Quadratic Pseudo-Boolean Optimization (QPBO) [3][4] is used to min-
 168 imize the proposed energy function, by merging a set of candidate matches for
 169 every superpixel in the first frame. For instance, for a given superpixel in the
 170 initial frame, the corresponding matching would be the most similar one in terms
 171 of color, shape, or the spatial distance. More candidate solutions can be added
 172 by defining a neighbourhood in the second frame and select random pairs from
 173 every neighbourhood of every superpixel in the first frame.

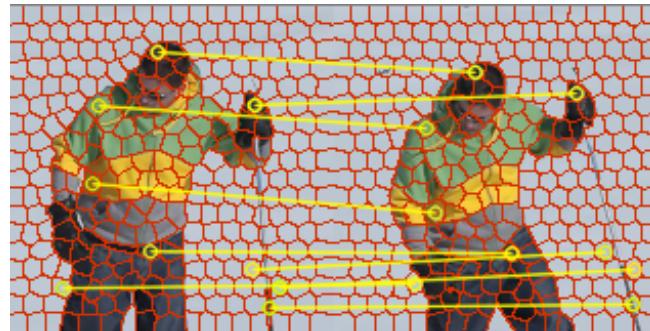


Fig. 2. The yellow lines show selected superpixel matching between a pair of distant frames in the Snow Shoes sequence.

The Fig. 2 shows results for large separations between frames. For this case, however, the matches in the textureless part of the scene are mostly invalids. Though this is expected because of the aperture problem and heavy occlusions.

2.2 Background regions tracking for object segmentation

The main idea to perform object segmentation consist in tracking (or more exactly, match) superpixels that are labeled as background, thanks to an object tracker initialization. Thus, the superpixels that are initially outside the tracker region of interest, can be propagated through the sequence, and if they fall into the window on a subsequent frame, they can be safely labeled as background (Fig. 3).

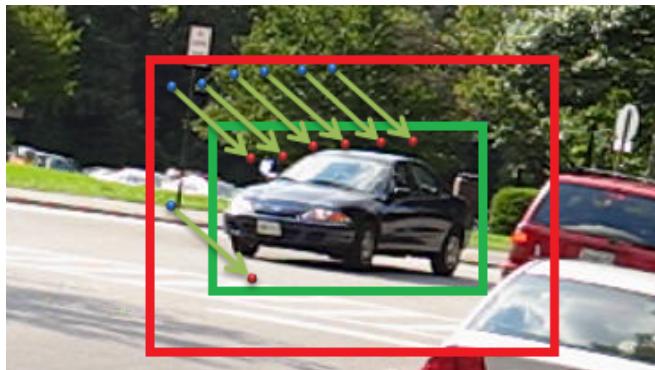


Fig. 3. Example image of points entering a tracking region (green) due to object motion in a video sequence.

To save computational power, the tracked superpixels are limited to the ones that fall inside a control region (red box in the Fig. 3). Usually, after several frames, the labeled superpixels will almost completely cover the unwanted areas in a dinamyc scene. We call this process background segments tracking. The Fig. 4 shows this idea in a real scenario. From left to right, initially the superpixels with elements outside the bounding box are labeled as background (green), then, as the sequence changes, the labeled superpixels flow inside the window, giving hints for the model initialization in the background-foreground separation algorithm. At this point, some generic segmentation technique can be connected to the pipeline to refine the segmentation (e.g. region growing). We prefer, however graph based segmentation methods ([18][15]) because the usual user interaction can be replaced by the tracked background regions.

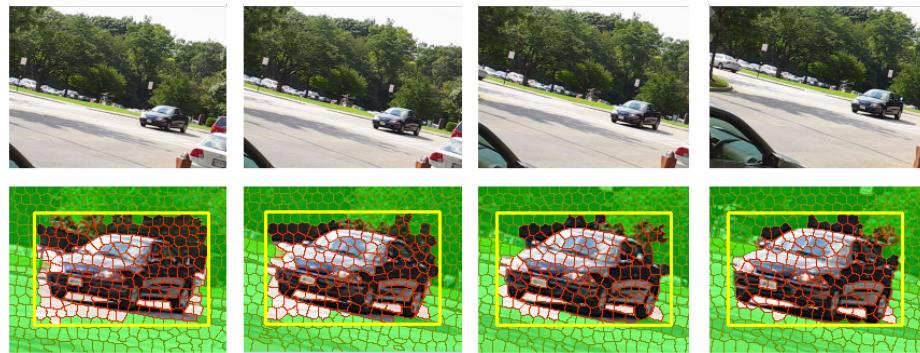


Fig. 4. Background segments automatic labeling and propagation, the flow goes from left to right.

2.3 Segmentation results

Fig. 5 shows the results for an image sequence where the interest object is the head of a person. The head tracker and the superpixel flow provide information for better background-foreground separation. The background-foreground models are updated as the frames go on, giving more robustness for sequential propagation of the segmentation. The method is tested in the Walking Couple sequence, by allowing only a small amount of iterations in the graph based segmentation. Observe how the contour in the man's head is correctly delineated when another person's head occludes part of it. In this case, the superpixels that belong to the womans face were correctly propagated and thus, labeled as background.

In order to understand the effect of including superpixel propagation in a video sequence for object segmentation, some results are shown in the Fig. 6. For these experiments only one iteration is allowed in the graph-cut based methods.

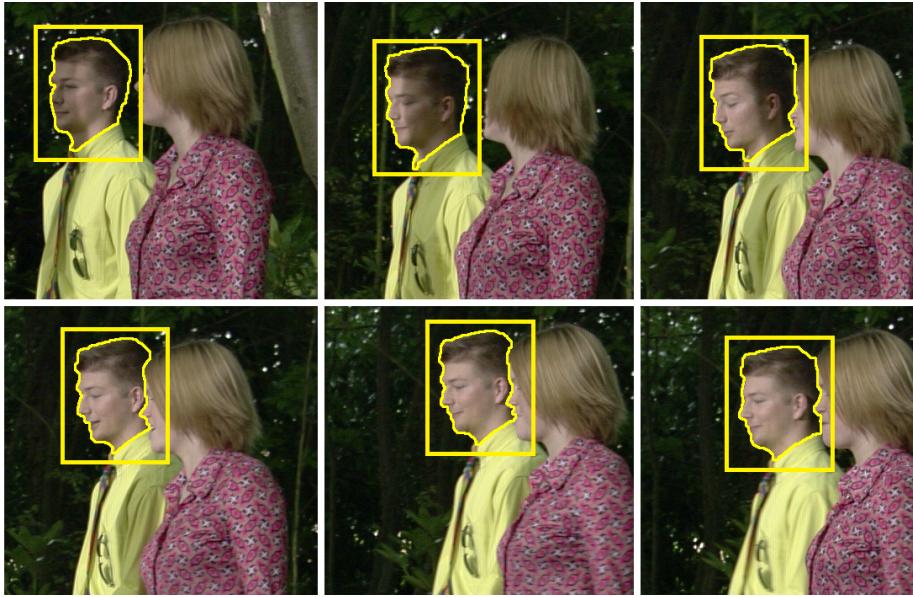


Fig. 5. Segmentation through the sequence Walking Couple (Yellow contour) initialized in the mans head. The yellow box correspond to the tracker output. The labeled background superpixel are not shown for clarity.

The top row frames (Fig. 6) were initialized only with the tracker, and the bottom row was initialized with the superpixel tracking technique. Observe that in general, the contour delineated is usually better in terms of precision and stability for the later one.

2.4 Flow estimation

The object flow consist on computing the motion field for an object of interest through an image sequence. The most usual approach to solve a problem like this is to implement some of the available optical flow techniques through the complete sequence and perform the flow integration. However, this process results in high levels of motion drift [18][19] and usually the motion of the interest object is affected by a global regularization. In some extreme cases, the interest object motion may be totally blurred and other techniques have to be incorporated. Moreover, the diversity of natural video sequences makes difficult the choice of one technique over another, even when specialized databases are at hand [17], because currently no single method can achieve a strong performance in every of the available datasets. Most of these methods consist in the minimization of an energy function with two terms (As was previously mentioned in the Sec. 2.1). The data term is mostly shared between different approaches, but the prior or spatial term is different, and basically states under what conditions the optical flow smoothness should be maintained or not. In a global approach,

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

356

357

358

359

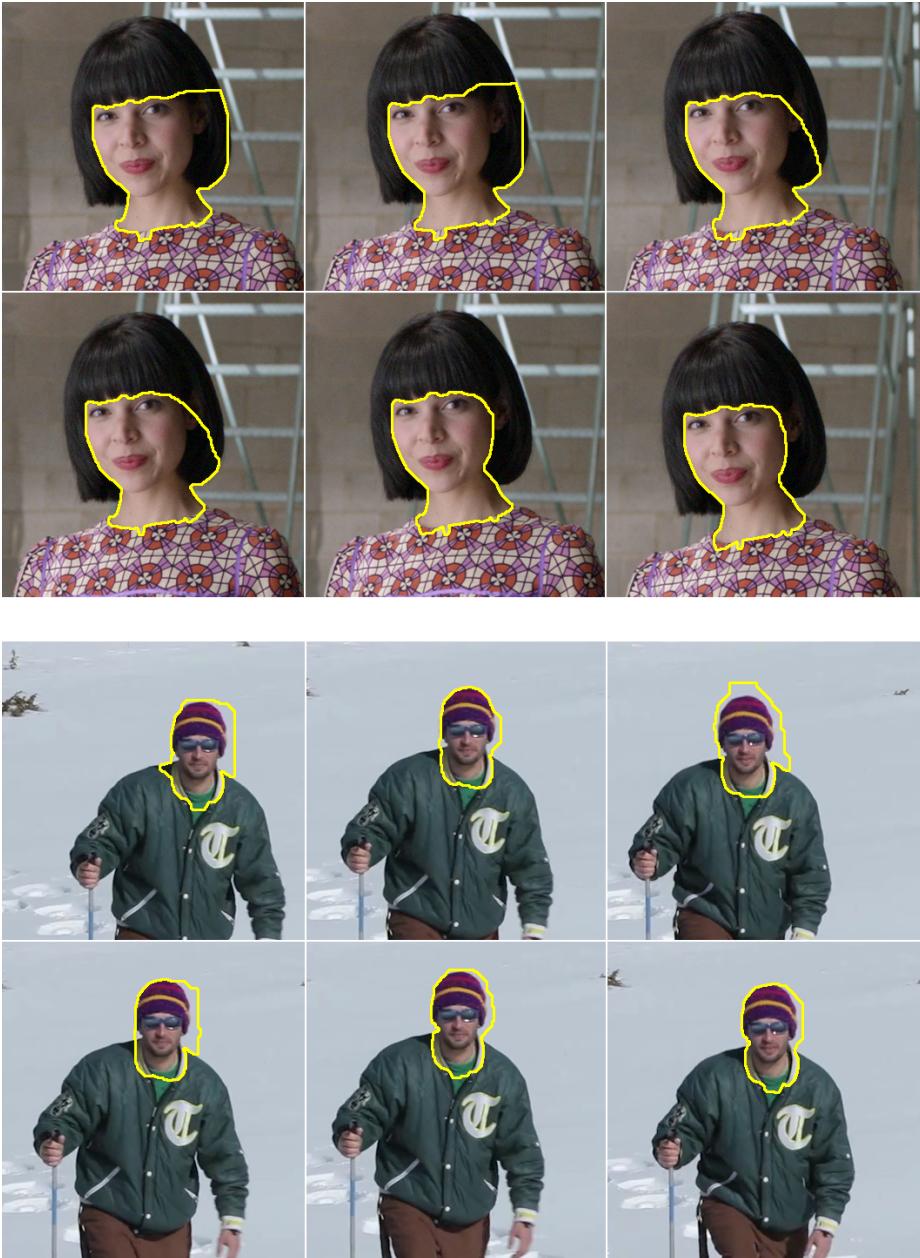


Fig. 6. Face segmentation in the Amelie Retro and the Snow shoes sequences in three different frames. For each group, the Top Row: One-iteration window-based graph-cuts; and the Bottom Row: One-iteration graph-cuts initialized with superpixel tracking.

315

316

317

318

319

320

321

322

323

324

325

326

327

328

329

330

331

332

333

334

335

336

337

338

339

340

341

342

343

344

345

346

347

348

349

350

351

352

353

354

355

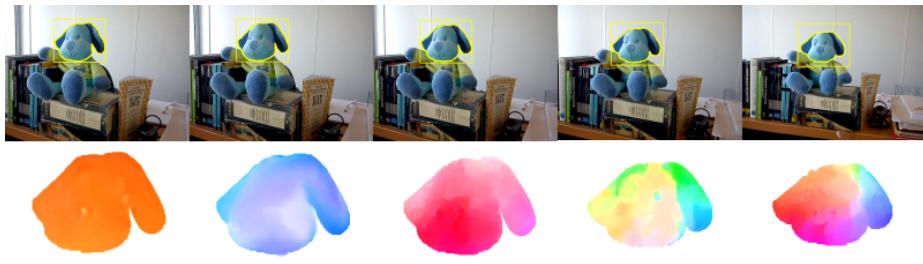
356

357

358

359

360 however, this is a difficult concept to define. Most of these smoothness terms
 361 rely in appearance differences or gradients. All these meaning that, unavoidably,
 362 some methods may be more reliable for some cases but weaker for others. It can
 363 be argued that this behaviour may be caused because most of the techniques do
 364 not count with a way to identify firmly where exactly this smoothness prior can
 365 be applied. The main idea behind the object flow is that given the availability



368
 369
 370
 371
 372
Fig. 7. Object flow with the color code of [17] (bottom)
 373 for frames in the Puppy
 374 sequence (up).
 375
 376

377
 378
 379
 380
 381 of several robust tracking techniques, and the proposed segmentation method
 382 for video, the optical flow computation can be refined by computing it succes-
 383 sively between pairs of tracked windows. The basic proposal to perform this
 384 refinement consist on considering the segmentation limits as reliable smoothness
 385 boundaries. This is, of course, under the assumption that the motion is indeed
 386 smooth within the object region. This is assumption is not far from reality in
 387 most scenes with an interest object. Naturally, as the object tracker is included,
 388 is expected that the object flow should be more robust to rapid motions than the
 389 optical flow. Thus, the full motion is split in two, the long range motion, given
 390 by the tracker window, and the precision part, given by the targeted optical flow.
 391 The Fig. 7 shows the object flow for a frame in the Puppy sequence. Observe
 392 the motion vectors are computed only inside the object of interest, preserving a
 393 strong smoothing prior, but also allowing internal variations in the flow.

394 As a first approximation to the object flow, the Simple Flow technique [21]
 395 is taken as core base. This is because of its scalability to higher resolutions and
 396 because its specialization to the concept of object flow is only natural. This is
 397 because in the Simple Flow pipeline the smoothness localization can be easily
 398 specified through computation masks. More specifically, the initial computa-
 399 tion mask is derived from the segmentation performed as prior step. The resulting
 400 flow is then filtered only inside the mask limits to enhance precision and fas-
 401 tening the implementation. However, direct modifications in other optical flow
 402 methods can be further studied. For instance, in graph-cut based minimization
 403 approaches, the regularity constraints can be precisely targeted by disconnecting
 404 foreground pixels from background ones.

405

3 Experimental results

406

407

408

409

410

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431

432

433

434

435

436

437

438

439

440

441

442

443

444

445

446

447

448

449

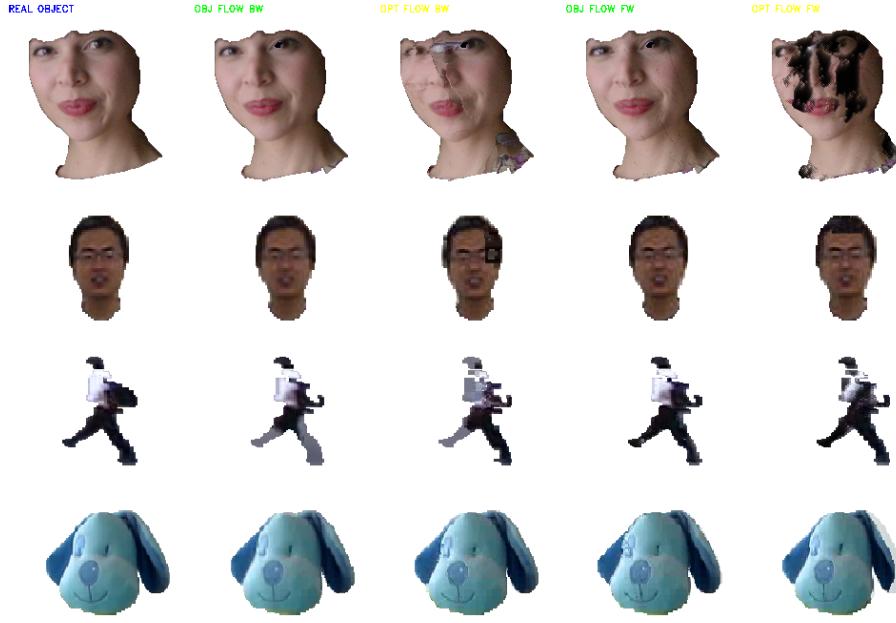


Fig. 8. Extrapolation results from integrated flow in 4 sequences. In descending order: Amelia Retro, Boy, Walking, Puppy. From Left to Right: Annotated object, Backward object flow, Backward optical flow, Forward object flow, Backward optical flow.

To evaluate the performance of the object flow in comparison with optical flow techniques, we performed a number of experiments on several video sequences. We annotated an initial bounding box for the videos, and a segmentation contour of the interest object for every frame. The experiment measures the ability of the method to extrapolate an image from the initial frame and the integrated flow. For every pair of frames in the video sequence, the PSNR between the annotated current state of the object and the extrapolated images is computed. The Fig. 8 is a sample of the performed experiment, each column is an image generated from the given flow. Two types integration are evaluated, *From – the – reference*, or forward integration, and *To – the – reference*, or backward integration, as discussed in [20]. So, for each row in the Fig. 8, two columns correspond to the object flow, and two columns correspond to optical flow, with both types of integration.

The Fig. 9 shows PSNR graphics for 4 different sequences. For every pair of frames an image is extrapolated, and the PSNR with the ground-truth object is computed. The results are shown with both, Euler integration (Labelled as *forward* in the figs.) of the used flow, and using the integration method described

450 in [20], labeled as *backward* in the figures. The results show that the object flow
 451 methods are generally more precise than its optical flow counterparts. Moreover,
 452 the object flow method with backward integration usually performs much better
 453 than any other combination of techniques. For this experiment, the object flow
 454 is compared with the simple-flow optical-flow method.
 455

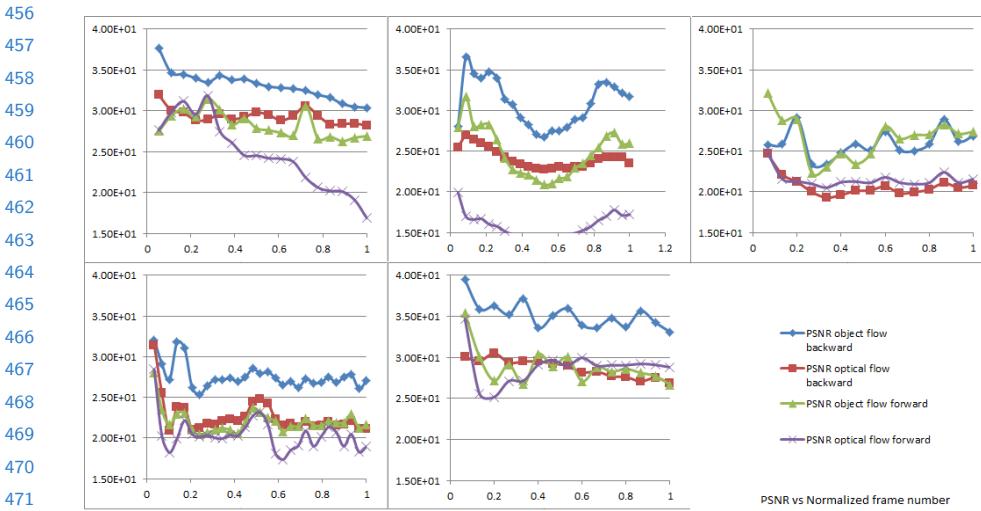


Fig. 9. PSNR graphs for extrapolated images using Object flow and the Simple Optical Flow for 4 sequences. In descendent order: Puppy Seq.; Amelie Retro Seq.; Boy Seq.; Walking Seq.

The Fig. 10 presents a visual comparison between the object flow and several optical flow techniques in the Amelia sequence for object extrapolation, and the involved frames (the first and last used frames in the sequence). The Fig. 11 shows the PSNR results for every extrapolated frame in the full sequence, the object flow performs better than all the studied optical flow techniques.

Observe that the object details are lost in comparison with the ground-truth object image (Fig. 10). For example, the closed eyes detail is missing in the most of the optical flow methods. Furthermore, several of the methods lost any significance, and the output barely holds any resemblance with the original image.

4 Conclusions

A framework to combine tracking and optical flow methods to improve object based dense motion description is presented. The pipeline is composed of three main steps, object tracking, segmentation and flow estimation. For the segmentation step a new promising video object segmentation algorithm was proposed, and, to the best of our knowledge, the introduced superpixel flow is the first energy based algorithm for superpixel matching. For the last step, we presented



Fig. 10. Top: Comparison between extrapolated objects using several methods: Groundtruth object, Object flow, TVL1, Block Matching, Brox, Farneback and Simple Flow. Bottom: First and current frame. The extrapolation is performed using backward accumulation of the flow.

a flow estimation method based on a modification of the simple-flow method to use the obtained segmentation mask. The experiments showed that this object based flow estimation improves the dense motion estimation in comparison to optical flow techniques. Future work can be further explored in the use of the object flow as feedback hint for tracking-by-detection methods. Also, several kind of applications of the object flow can be more deeply approached. For instance, in the structure-from-motion pipeline, video based rendering, automatic video edition, and video inpainting among others.

References

1. J. Malik and X. Ren, Learning a classification model for segmentation, *Computer Vision, International Conference*, 2003.
2. S. Boltz; F. Nielsen and S. Soatto, Earth mover distance on superpixels *International Conference on Image Processing*, 2010.
3. E. Boros; P. Hammer and G. Tavares, Preprocessing of unconstrained quadratic binary optimization *RUTCOR*, 2010
4. E. Boros and P. Hammer, Pseudo-boolean optimization *Discrete applied Mathematics*, 2002
5. B. Horn and B. Schunck, Determining Optical Flow *Artificial Intelligence*, 1981
6. H. Ishikawa and P. Boualem, Multimodal estimation of discontinuous optical flow using Markov random fields *TPAMI*, 1993

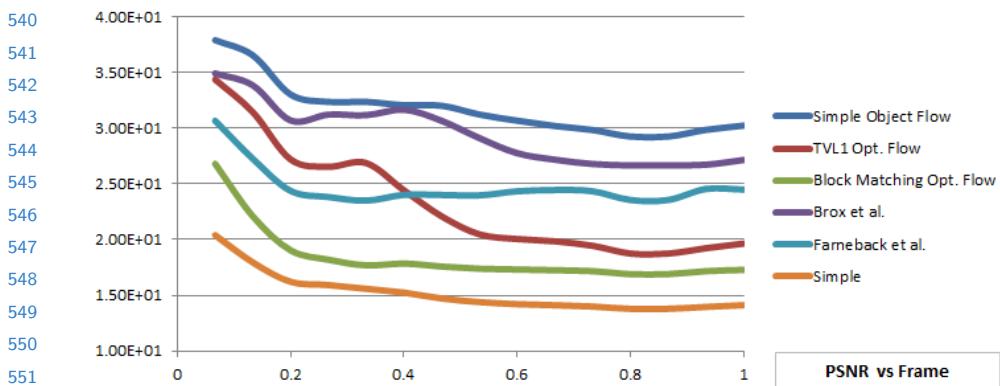


Fig. 11. PSNR graphs for extrapolated images using Object flow and the different Optical Flow techniques for the Amelia sequence.

- 540
541
542
543
544
545
546
547
548
549
550
551
552
553
554
555
556
557
558
559
560
561
562
563
564
565
566
567
568
569
570
571
572
573
574
575
576
577
578
579
580
581
582
583
584
7. V. Lempitsky, S. Roth and C. Rother, Fusion Flow: Discrete-Continuos optimization for optical flow estimation *Computer Vision and Pattern Recognition*, 2008
 8. M. Reso and J. Jachalsky, Temporally Consistent Superpixels *International Conference Computer Vision.*, 2011
 9. R. Achanta; A. Shaji; K. Smith; Aurelien Lucchi; P. Fua and S. Susstrunk SLIC Superpixels compared to state of the art superpixel methods *Discrete applied Mathematics.*, 2002
 10. F. Perbet and A. Maki, Homogeneous superpixels from random walks *MVA.*, 2011
 11. C. Xu and J.J. Corso. Evaluation of super-voxel methods for early video proccesing. *Computer Vision and Pattern Recognition*. 2012.
 12. A. Shekhovtsov, I. Kovtun and V. Hlavac. Efficient MRF deformation model for non-rigid image matching. *Computer Vision and Pattern Recognition*. 2007.
 13. J. Sun, N.N Shen and H.Y. Shum. Stereo matching using propagation belief. *TPAMI*. 2003.
 14. C. Rother, V. Kolmogorov and A. Blake. Grabcut: Interactive foreground extraction using iterated graph cuts. *SIGGRAPH*. 2004.
 15. L. Yang, Y. Guo, X. Wu and X. Wang. A new video segmentation approach: Grabcut in local window. *Soft Computing and Pattern Recognition*. 2011.
 16. Y. Wu, J. Lim and M.-H. Yang. Online object tracking: A benchmark. *Computer Vision and Pattern Recognition*. 2013.
 17. S. Baker, D. Scharstein, J.P. Lewis, S. Roth, M.J. Black and R. Szeliski. A Database and Evaluation Methodology for Optical Flow *International Journal Computer Vision*. 2013.
 18. Y. Boykov, M-P. Jolly. Interactive Graph Cuts for Optimal Boundary & Region Segmentation of Objects in N-D images *International Conference of Computer Vision*. 2013.
 19. W. Li, D. Cosker and M. Brown. An anchor patch based optimization framework for reducing optical flow drift in long image sequences. *Asian Conference of Computer Vision*. 2012.
 20. T. Crivelli, P.-H. Conze, P. Robert, M. Fradet and P. Perez. Multi-step flow fusion: towards accurate and dense correpondence in long video shots. *British Conference Machine Vision*. 2012.

CONFIDENTIAL REVIEW COPY. DO NOT DISTRIBUTE.

14 ACCV-14 submission ID ***

- 585 21. M. Tao, J. Bai, P. Kohli, and S. Paris. SimpleFlow: A Non-iterative, Sublinear
586 Optical Flow Algorithm. *Computer Graphics Forum, Eurographics*. 2012.
587 22. S. Hare, A. Saffari, P.H.S. Torr. Struck: Structured output tracking with kernels.
588 *International Conference of Computer Vision* . 2011.
589 23. B. Babenko, M.-H. Yang and S. Belongie. Robust Object Tracking with Online
590 Multiple Instance Learning. *Pattern analysis and Machine Learning* . 2010.
591
592
593
594
595
596
597
598
599
600
601
602
603
604
605
606
607
608
609
610
611
612
613
614
615
616
617
618
619
620
621
622
623
624
625
626
627
628
629

585

586

587

588

589

590

591

592

593

594

595

596

597

598

599

600

601

602

603

604

605

606

607

608

609

610

611

612

613

614

615

616

617

618

619

620

621

622

623

624

625

626

627

628

629