Object flow

Anonymous ACCV 2012 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 super pixel 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 scenarions 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 motion problem remains a challenge. However, the problem is more bearable for object tracking techniques. Moreover, even when object trackers and optical flow give good hints for object segmentation in video, these 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 copes with segmentation of the target.

Among the state of the art segmentation methods for objects in video sequences, point trajectories based ones stand for its performance and reliability, even when only sparse trajectories are known because of computational reasons. In the other hand, for the problem of extracting out a preselected object in still frames, max-flow min-cut based approaches have demostrated to be a powerful tool. We propose to mix these two ideas together with the tracking of backgorund regions via the novel concept of superpixel flow for reliable object segmentation through video. We show how this extra information can be used to complement the graph-cuts based techniques for an efficient foregorund-background segmentation.

The present paper is organized as follows. We introduce the concept of superpixel flow in Sec. 2. Then, a method for object segmentation in video which uses object tracker information and the background segments tracking is presented in Sec. 3. In following sections the object flow basic approach is explained. Finally, results showing how the object flow overpass state of the art optical flow methods for object motion flows computation.

2 Superpixel flow

2.1 Problem definition

Superpixels and over segmentation techniques became a widely used pre-processing stage for a large number of machine vision applications, after the original concept was introduced [1]. Superpixels are traditionally used as performance booster for several other techniques. However, it is still mostly related to single frame processing [1][10][11]. In the search for consistency in superpixel labeling through video, some authors have proposed different techniques, which go from simple extension to supervoxels[9][11], to more complicated approaches [8]. These approaches, nonetheless, usually require a global processing and knowledge of all (or several of) the video frames beforehand. We propose a superpixel matching technique which assumes a flowlike behavior in the image sequences (natural video). Some previous work have been done towards a superpixel based image comparison using the Earth Mover's Distance, by taking superpixels as bins of a global histogram [2]. The label propagation or superpixel flow can be achieved with this technique as a byproduct, by selecting the superpixel in the second frame that maximize the flow from each superpixel in the first frame.

By taking into account superpixels computed separately in images, so the video process can be performed with only two frames at a time, we move towards a more time efficient approach. 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). In this sense, it seems natural that the problem of

superpixel flow could be solved with a discrete energy minimization procedure. 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 $l_p \in 0, 1, ...N-1$) between a pair of frames (I_0,I_1), but holding a flow-like behavior.

Thus, the superpixelization should maintain certain size homegenity 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 homegenity of the superpixelization across the sequence. The proposed steps to solve the propagation problem assume this requirement is hold. For other kind of the techniques, other approaches should be followed.

2.2 Energy Formulation

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

$$P(l|I_0, I_1) = \prod_{p \in \Omega} \mathrm{e}^{-D_p(l_p; I_0, I_1)} \prod_{p,q \in \mathcal{N}} \mathrm{e}^{-S_{p,q}(L_p; L_q)} \tag{1}$$

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

$$E(l) = \sum_{p \in \Omega} D_p(L_p; I_0, I_1) + \sum_{p,q \in \mathcal{N}} S_{p,q}(L_p, L_q) \tag{2}$$

The terms D, and S in 2 stand for data term and spatial smoothness terms as they are popularly known in the MRF literature. The first one determines how accurate is the labeling in terms of consistency of the measured data (color, shape,etc.). In the classical optical flow formulation of this equation, the data term corresponds to the pixel brightness conservation[7][5]. However, as superpixels are a set of similar (or somehow homogenous) pixels, an adequate color based feature can be a low binned histogram or its average color. So it can be written more precisely as

$$D_p(l_p; I_0, I_1) = \rho(h(p), h(p')) \tag{3}$$

Where $h(p)$ and $h(p')$ are the histograms of the superpixel p and its correspondent superpixel in the second frame (I_1). The distance ρ can be replaced

by the Bhattacharyya distance. Note that the low binned histogram or average color gives certain robustness against noise, and slowly changing colors between frames. The spatial term is a penalty function for horizontal and vertical changes of the vectors that have origin in the centroid of the superpixel of the first frame and end in the centroid of the super-pixel of the second frame.

$$S_{p,q}(l_p, l_q) = \lambda(p)\sqrt{\frac{|u_{p_c} - u_{q_c}|}{\|p_c - q_c\|} + \frac{|v_{p_c} - v_{q_c}|}{\|p_c - q_c\|}} \tag{4}$$

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

In 4 the operator ρ has the same meaning as in the data term 3. The histogram distance is nonetheless computed between superpixels p and q, which belong to the same neighborhood. The superpixels centroids are noted as q_c and p_c, and u and v are the horizontal and vertical changes between centroids. This term is usual in the MRF formulation and has a smoothing effect in superpixels that belong to the same object. It has to be observed that when two close superpixels are different, thus, more probable to belong to different objects within the image, the term λ allows them to have matches that do not hold the smoothness prior with the same strength. It has to be noted that the proposed energy function is highly non-convex.

2.3 Energy Minimization

A fair amount of work had been dedicated to discrete optimization techniques in computer vision, leading to a couple of well-defined and widely tested approaches to solve the pairwise MRF[3][4]. However, some of the approaches restrict the construction of the spatial term, and/or enforce limitations in the number of labels [3]. Because of the high amount of possible labels for each element in the proposed approach, the use of the Fusion Moves [7] technique seems to be well suited. This algorithm employs the Quadratic Pseudo-Boolean Optimization (QPBO), to combine incremental sets of proposal labelings, resulting a semi-globally-optimal solution [4]. Thus, the minimization starts by proposing a set of possible solutions, and iteratively merge them with the QPBO technique. The possible solutions that can be given depend on the kind of problem that is intended to be solved. For example, in stereo superpixel matching, some assumptions related to the cameras organization can be made to generate solutions. In a more generic sense, other assumptions can be made towards option generation. For instance, for a given superpixel in the initial frame, in the second frame the corresponding matching would be the most similar one in terms of color, or the most similar in terms of shape, or the spatially closer superpixel. More proposal solutions can be added by defining a neighborhood in the second frame and select random pairs from every neighborhood of every super-pixel in the first frame. This is more suitable for problems where the images are extracted from the same video sequence. To speed-up the minimization procedure, the QBPO properties can be exploted. For instance, the fusion of the proposed solutions is

always guaranteed of lowest or equal energy than the two proposals. Thus, one could split the fusion procedure in several cores and build a hierarchichal chain as fusions of proposal are subsequently fused.

2.4 Matching results

The Fig. 1 shows some examples of superpixel matching with the presented method. It can be seen that the matching performs well even in difficult cases, like the hands in the top row. It has to be noted as well that even in superpixels where there is a lack of texture, there is correct matching. This seems to be the effect of enforcing the regularization between superpixels that are close, but are also similar to each other.

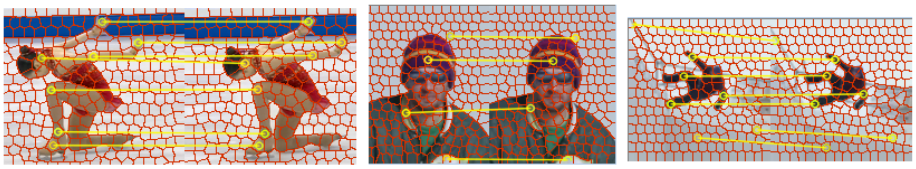

Fig. 1. The yellow lines show selected superpixel matching between pairs of consecutive frames in a video with the proposed method. The video frames go from right to left.

Moreover, unlike most of the optical flow methods, superpixel flow extends naturally for more distant frames. The Fig. 2 shows results for larger separations between frames, without tweaking or adjusting any parameters. For this case, however, the matches in the textureless part of the scene are mostly invalids. Though this is expected because because of the aperture problem and heavy occlusions.

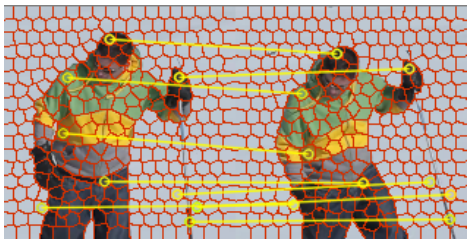

Fig. 2. The yellow lines show selected superpixel matching between a pair of distant frames in the Snow Shoes sequence.

3 Superpixel propagation for object segmentation in videos

The algorithm proposed in [18], offers a good deal in terms of background-foreground separation from user interaction. A technique like this, however, performs very well in still images, but it may not be well adapted for sequential videos, Extensions to this method, like, GrabCut ([14]) work by implementing an iterative graph cut based minimization to separate regions according to appearance information, that can be extracted from the user interaction. This interaction, however, could be minimized in videos, given the extra information that offers the flow of the sequence. Some authors had approached the GrabCut or similar graph based segmentation techniques in sequential videos, to propagate a consistent segmentation [15]. However, some more work on reducing user interaction given the extra flow-like information that video sequences offer is still needed. We propose to combine the presented superpixel flow as an automatic method to initialize the desired min-cut max flow based algorithm to perform object segmentation through frames in a video sequence.

The main idea 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 ROI, can be propagated through the sequence, and if they fall into the ROI of the next frame, they can be safely labeled as background again. We call this process background segments tarcking. The process is repeated for any labeled superpixel through the video. Having several labeled superpixels can reduce widely or totally the necessity for user interaction in subsequent frames. Thus, to perform object segmentation in a full video sequence, the required user interaction would only be the initial bounding box. Moreover, a fully automatic approach can be obtained if a reliable object detector is available.

Fig. 3 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. 4. For these experiments only one iteration is allowed in two grab-cut based methods. One initialized only with the tracker, and the other complemented with the superpixel propagation. Observe that in general, the contour delineated is usually better in terms of precision and stability for the later one.

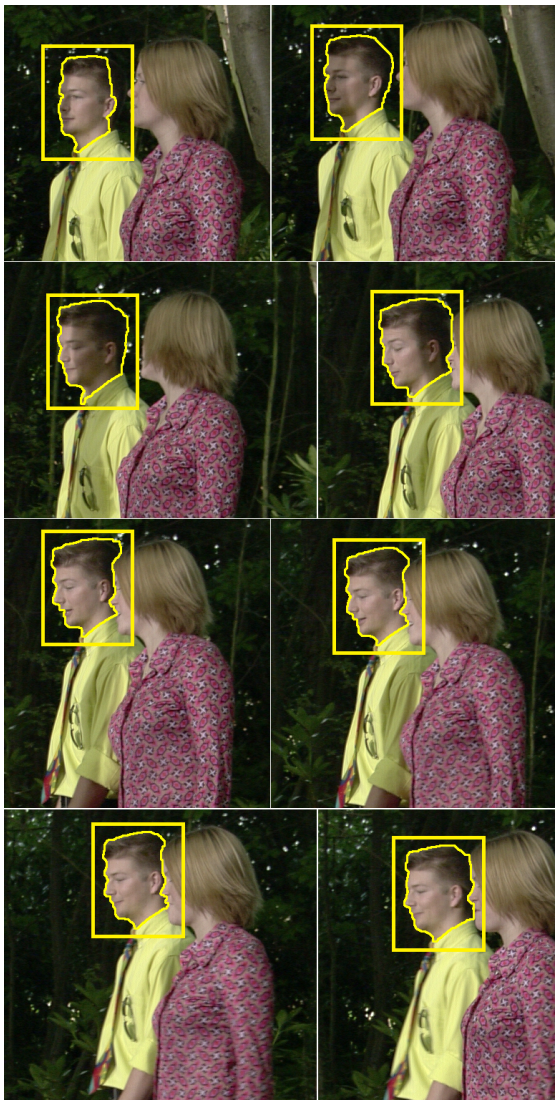

Fig. 3. Segmentation through the sequence Walking Couple (Yellow contour) initialized in the mans head. The yellow box correspond to the tracker output.

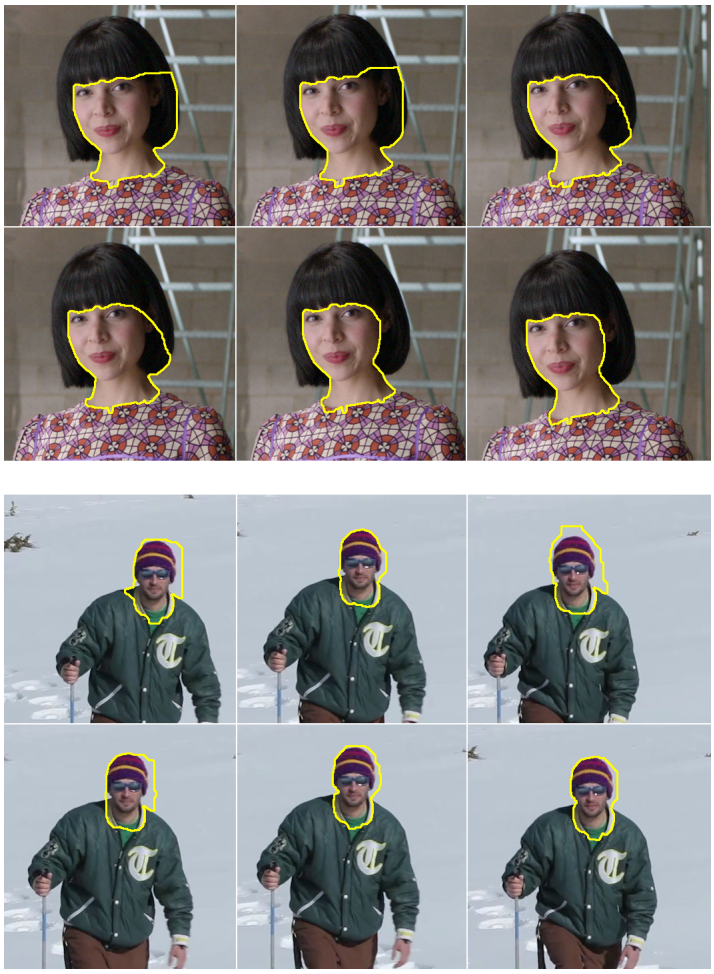

Fig. 4. Face segmentation in the Amelie Retro and the Snow shoes sequences in three different frames. For each group, the Top Row: One-iteration window-based grabcut; and the Bottom Row: One-iteration grabcut with super pixel propagation.

4 Object flow

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). 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, however, this is a difficult concept to define. Most of these smothness terms rely in appearance differences or gradients. All these meaning that, unavoidably, some methods may be more reliable for some cases but weaker for others. It can be argued that this behaviour may be caused because most of the techniques do not count with a way to identify firmly where exactly this smoothness prior can be applied. The

Fig. 5. Object flow with the color code of [17] (Right) for one frame in the Puppy sequence (Left). Green Box: Current tracker state. Yellow: Segmentation contour.

main idea behid the object flow is that given the availability of several robust tracking techniques, and the proposed segmentation method for video, the opti-

cal flow computation can be refined by computing it successively between pairs of tracked windows. The basic proposal to perform this refinement consist on considering the segmentation limits as reliable smoothness boundaries. This is, of course, under the assumption that the motion is indeed smooth within the object region. This is assumption is not far from reality in most scenes with an interest object. Of course, as the object tracker is included, is expected that the object flow should be more robust to rapid motion than the optical flow. Thus, the full motion is split in two, the long range motion, given by the tracker window, and the precision part, given by the targeted optical flow. The Fig. 5 shows the object flow for a frame in the Puppy sequence. Observe the motion vectors are computed only inside the object of interest, preserving a strong smoothing prior, but also allowing internal variations in the flow.

4.1 Implementation details and results

As a first approximation to the object flow, the Simple Flow technique [21] is taken as core base. This is because to its scalability to higher resolutions and because its specialization to the concept of object flow is only natural. This is because in the Simple Flow pipeline the smoothness localization can easily especified trough computation masks. More specifically, the initial computation mask is derived from the segmentation performed as prior step. The resulting flow is then filtered only inside mask limits to enhance precision and fastening the implementation. However, direct modifications in other optical flow methods can be further studied. For instance, in graph-cut based minimization approaches, the regularity constraints can be precisely targeted by disconecting foreground pixels from background ones.

Fig. 6. Extrapolation results from integrated flow in one frame of the Amelie Retro sequence. From Left to Right: Anotated 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 anotated 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 the PSNR between the anotated

current state of the object and the extrapolated images is computed. The Fig. 6 is a sample of the performed experiment, each column is an image generated from the given flow.

The Fig. 7 shows PSNR graphics for 4 different sequences. For every pair of frames an image is extrapolated, and the PSNR is computed. The measure is computed using Euler integration (Labeled as *forward* in the figs.) of the used flow (object or optical flows), and using the integration method described in [20], labeled as *backward* in the figures.

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. Bouthemy, Multimodal estimation of discontinous optical flow using Markov random fields*TPAMI.*, 1993
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, Homogeneus 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 on Computer Vision*. 2013.

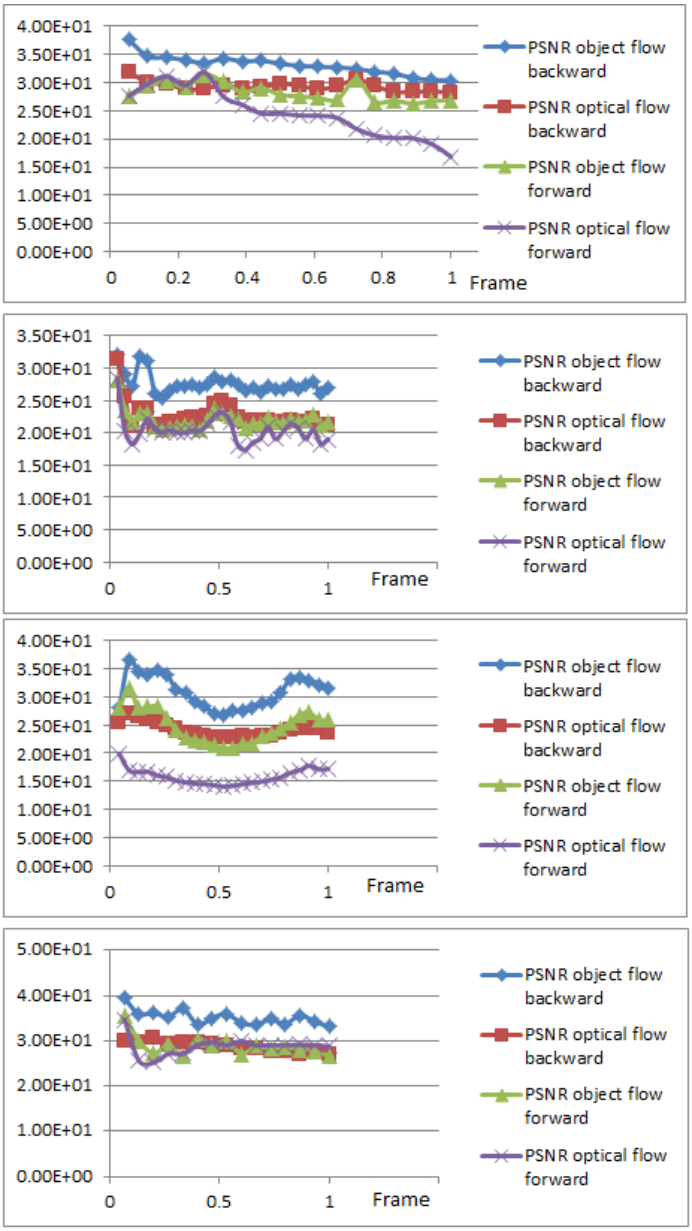

Fig. 7. PSNR graphs for extrapolated images using Object flow and the Simple Optical Flow for 4 sequences. In descendent order: Puppy Seq.; Amelie Retro Seq.; Boy Seq.; Walking Seq.

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 on 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.
21. M. Tao, J. Bai, P. Kohli, and S. Paris. SimpleFlow: A Non-iterative, Sublinear Optical Flow Algorithm. *Computer Graphics Forum, Eurographics*. 2012.