Proceedings of the World Congress on Electrical Engineering and Computer Systems and Science (EECSS 2015)
Barcelona, Spain, July 13 - 14, 2015
Paper No. 329

Motion Segmentation with Reduced Prior Knowledge Dependency

Hodjat Rahmati, Ole Morten Aamo, Øyvind Stavdahl
Cybernetics Engineering/ NTNU
O. S. Bragstads plass 2D, No7491, Trondheim, Norway
{hodjat.rahmati, ole.morten.aamo, oyvind.stavdahl}@itk.ntnu.no

Ralf Dragon
Computer Vision Laboratory/ ETH
Sternwartstrasse 7, CH - 8092 Zürich, Switzerland
dragonr@vision.ee.ethz.ch

Abstract- Segmentation and tracking of objects with fast and complicated motion patterns is a difficult task in computer vision, particularly in presence of occlusions. Motion information is a strong cue to distinguish objects from each other in a video, the more different patterns two objects have the easier separating them will be. In situations where different objects show similar movements, motion information is not informative enough by itself. In such cases, prior knowledge can complement the motion information. A way of providing prior knowledge is to manually label some trajectories. Providing enough manual labels may be straining, especially when segmentation is based on trajectories that terminate at some point in time, requiring additional manual labeling. To decrease the need for prior knowledge, we propose a new particle matching technique that employs multi-scale optical flow in order to re-detect particles through the video shot. Experimental results demonstrate the reliability of our method.

Keywords: Motion Segmentation, Particle Matching, Tracking, Cerebral Palsy Prediction

1 Introduction

Motion segmentation can be defined as grouping point trajectories over image sequences. While earlier work focused on assigning the trajectories to subspaces, e.g. with the generalized PCA [13], subsequent work exploited sparsity [5] or non-negative matrix factorization [2]. Further works exploit temporal smoothness [10] or depth ordering [4]. In the most recent works, [1, 3, 5], the pairwise relationships between trajectories are aggregated and a final spectral clustering finds the association of trajectories to motion segments.

The aforementioned methods are all unsupervised. In case of fast and complicated motion patterns unsupervised methods fail to perform reliably. Recently, we proposed a semi-supervised framework for integrating prior knowledge into their energy-minimization based motion segmentation approach [9]. The prior knowledge consists of a manual labeling to select a sample of each segment that represents that segment through the optimization process. Unlike image segmentation where only one label is needed for each segment, a cumbersome problem with trajectory segmentation is that the initially labeled trajectories might not last for the whole shot. This happens in case of occlusions or a fast motion. So all trajectories of a segment may end and, consequently, there is no trajectory left to represent that segment. The energy of assigning trajectories that are initialized from that moment on to that segment is very large, while it's smaller for the other segments [9]. It is more probable that these trajectories would not be labeled as belonging to

the terminated segment although it might be the right one. To overcome this problem [9] provides more manual labeling such that every 500th frames are manually labeled. In the present paper, we tackle this problem by equipping the method of [9] with a particle matching technique that automatically extends the prior knowledge.

The primary motive for our work is to extract the motion data out of videos with the necessary level of details to be used in predicting cerebral palsy (CP) in young infants. The method was used in [8] to predict cerebral palsy, with encouraging results.

2 Basic Motion Segmentation

In this section, we summarize the motion segmentation method proposed in [9] which is the base of our segmentation method. This method separates different objects from each other using their motion information and tracks them through the whole video shot. To do so, the first step is to build up a dense trajectory set tracking particles spread over the whole image. To distinguish objects, trajectories are segmented into different groups based on similarity in their motions. The task of splitting is performed by a graph-cut optimization. Each vertex of the graph represents a trajectory, and vertices are connected by weighted edges reflecting similarities between connected trajectories. This similarity is a measure of a combination of spatial and motion distances such that the closer two trajectories are and/or the more similar patterns their motions share, the higher their similarity will be. The output of this phase is a set of labels that assigns each trajectory to one of the pre-defined segments.

This motion segmentation method needs prior knowledge to base the segmentation upon. This prior knowledge is the assignments of a small subset of trajectories belonging to each segment. The true assignments of these trajectories are shared with the optimizer. Unlike image segmentation where the prior knowledge lasts during the optimization problem, a cumbersome problem with trajectory segmentation is that the initially labeled trajectories might not last for the whole shot. This happens in case of occlusions or a fast motions. To overcome this problem, in [9] we provided more prior knowledge by manually labeling trajectories in every 500th frames. In the following section, a new method is presented to overcome the problem as well as fulfilling our ultimate goal which is to perform motion segmentation with as little user interactions as possible. For simplicity we refer to our method in [9] as *basic moseg* and the proposed method as *improved moseg*.

3 Improved Motion Segmentation

Since trajectories belonging to an object are terminated due to fast motions or occlusion, we re-detect the object when reappearing. This is explained in Fig. 1 where a synthetic example is created to show the problem with *basic moseg* as well as the procedure in *improved moseg*. Each row represents a trajectory over time that starts at a frame and may end at another frame. The trajectories belong to two different segments, the shapes show the true assignments while the colors are the decisions of the motion segmentation method.

As can be seen, as long as there exist some manually labeled trajectories representing each of the segments at a frame, both methods end up with the correct assignments in that frame. However in frame M where the circle segment looses all of it's manually labeled trajectories, the segmentation is still correct. This is due to transitivity, trajectories that have common frames with trajectories that have common frames with the manually labeled ones could still get the correct assignment. The main problem with the *basic moseg* arises when not only all the manually labeled trajectories for a segment are terminated, but there is no trajectory left which has common frames with them, such as in case of complete occlusion. This happens in frame K where the *basic moseg* leads to a wrong segmentation. On the other hand, the *improved moseg* keeps up with the correct segmentation because it could find a particle in frame K that matches an initially labeled particle. In fact, its good performance is due to extending the manually labeled trajectories by finding new

ones that match them.

Our proposed method consists of two stages: first, objects are re-detected by finding matched particles, then new trajectories are initialized for the matched particles. In order to re-detect the objects, we obtain optical flow fields over multiple time scales. In other words, every kth frames is compared to the manually labeled frame and a flow field is calculated for each of them and finally matched points are found.

We use LDOF [11] to obtain the flow fields for two reasons. Firstly, it is designed for cases with large displacement, and secondly, it provides a very dense set of trajectories. Let's suppose that the flow field between the labeled frame and frame k is estimated. The next step is to derive matched points between these two frames. We initialize points in the labeled frame, and try to obtain matches for each of those points in frame k. For reducing the cost of matching and also because of unnecessity in matching points with no structure in their neighborhood, points are initialized in the areas with structure in their vicinity. To measure the structure, the structure tensor for a point in the image, $\mathbf{x}$, at each channel is obtained as follows,

$$\mathbf{S} = \mathbf{G} * \begin{bmatrix} \mathbf{I}_x^2 & \mathbf{I}_x\mathbf{I}_y \\ \mathbf{I}_x\mathbf{I}_y & \mathbf{I}_y^2 \end{bmatrix} \tag{1}$$

where $\mathbf{I}_x$ and $\mathbf{I}_y$ are the image derivatives in x and y directions, respectively, and $\mathbf{G}$ is a Gaussian kernel function with standard deviation $\sigma = 2$ centered on point with position $\mathbf{x}$. Then, the structure tensors of different channels are added up (here there are three channels, one for each color) to build the total structure tensor. Finally, points which second eigenvalue of their total structure tensor are smaller than a percent of average second eigenvalue for the whole image are removed from initialization.

The matches for initialized points are obtained by propagating each point in the labeled frame to frame k using the relevant forward optical flow field $\mathbf{w} := (u,v)^T$ via the following formula,

$$\hat{\mathbf{x}}_k = \mathbf{x} + \mathbf{w}. \tag{2}$$

Where $\hat{\mathbf{x}}_k$ is the estimated position of the match for $\mathbf{x}$ in frame k.

Due to occlusion and possibility of wrong motion estimation in the optical flow field, unreliable matches must be removed. To do so, forward and backward flow fields are compared to each other. There are large inconsistency between the forward and backward flow nearby the optical flow boundaries and occluding area [7]. If they are not consistent for an specific point, either that point is occluded or the estimated flow field is not reliable. In both cases the matched point is not reliable.

Let $\hat{\mathbf{w}} := (\hat{u},\hat{v})^T$ be the backward flow field, then for a non-occluded point with $\mathbf{w}_t(\mathbf{x}_t) = -\hat{\mathbf{w}}_t(\mathbf{x}_t + \mathbf{w}_t(\mathbf{x}_t))$. But because of inaccuracy in the flow estimation, we consider a tolerance bound such that if the difference of backward and forward flows exits this bound we ignore the match. So, as long as the following inequality is valid, the relevant match could be valid:

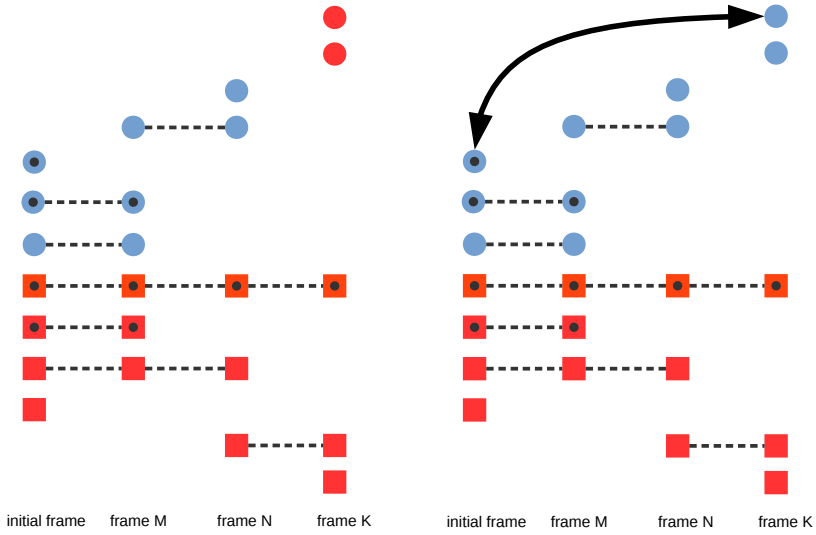
$$|\mathbf{w} + \hat{\mathbf{w}}|^2 < 0.01(|\mathbf{w}|^2 + |\hat{\mathbf{w}}|^2) + 0.5. \tag{3}$$

The tolerance bound is proportional to the motion size for the subjected point, and the larger the motion, the more error is acceptable. We also ignore matches on the motion boundaries to prevent drifting. Therefore matches with

$$|\nabla u|^2 + |\nabla v|^2 > 0.01|w|^2 + 0.002 \tag{4}$$

are deleted. This procedure is repeated for every kth frame from the manually labeled frame, and finally the match set is formed as the set of all these matches.

For the second stage, that is to establish new trajectories on the matched points, we employ the same procedure as for the initialized points on the first frame [11]. So, for each matched point on frame k a new



(a) *basic moseg* (b) *improved moseg*

Fig. 1: Segmentation for a synthetic example. A set of trajectories belonging to two segments are shown, each row represents a trajectory over time that starts at a frame and may end at another frame. The shapes show the true assignments while the colors are the decisions of the motion segmentation, the ideal segmentation leads to blue circles and red squares. Those with black circle in the middle are the manually labeled trajectories. The double-sided arrow shows the matched particles by *improved moseg*.

trajectory is created using the flow field of succeeding and preceding frames. Since the new trajectories are developed on a matched point, they get the same label as their matches on the manually labeled frame. Termination of initially labeled trajectories are compensated by creating these new trajectories, so the segments have some representative through the whole shot.

4 Experimental Results and Discussion

In all experiments, trajectories are developed as it is proposed in [11]. Two different data sets are used to study the performance of our proposed methods on. First, as the primary motivation for starting this work was to largely automate prediction of cerebral palsy, a set of infants' videos are studied in more details. Second, in order to investigate the applicability of our method, it is tested on a standard benchmark.

4.1 Performance on Videos of Infants

In experiments in this section, we used the experimental set-up of *St. Olavs Hospital*. During the experiments, we analyzed the first 1000 frames of 10 sequences of different infants carrying out different motions. These sequences are a magnitude longer than the Hopkins 155 [12] and the Freiburg-Berkeley [6] dataset with an average length of 30 and 245 frames, respectively. As ground truth, we manually annotated a dense segmentation of every 250th frame as displayed in Fig. 2. Due to occlusions, fast and complicated motion patterns, the trajectories last just for 96.5 frames in average.

The goal of segmentation in our application is to capture the motion of six different segments representing hands, feet, head and trunk. Fig. 3 shows the segmentation results for [1] which is an unsupervised technique. It is not hard to discover that this method could separate only very distinct motions from each other, and most of the trajectories are assigned to the background. Its poor performance can hardly be criticized. Fast and

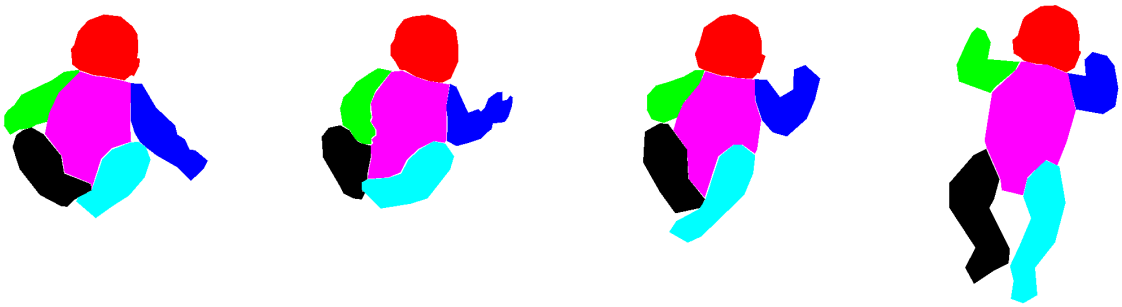


Fig. 2: Seq. 1 ground-truth segmentation for frames 1, 50, 200, 300 from left to right.

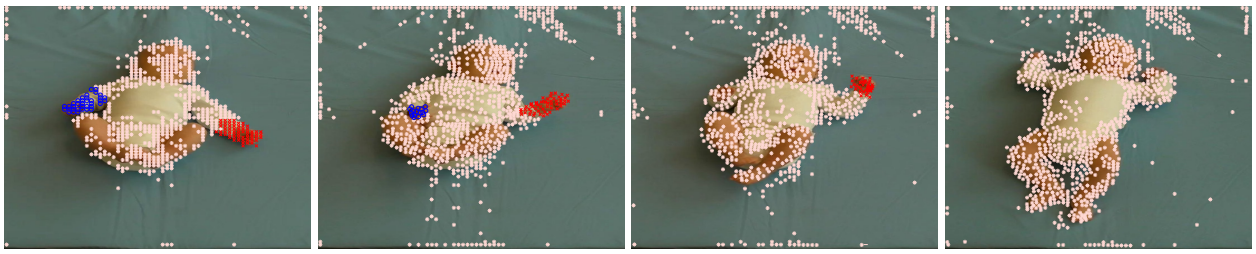


Fig. 3: Seq. 1 segmentation results of [1] for frames 1, 50, 200, 300 from left to right.

complicated motion patterns of body parts make the segmentation task challenging. Additionally, particles on the same desired segment move quite differently and without additional knowledge, motion is not a strong enough cue to meet the segmentation demands. This additional knowledge could be provided as a set of prior knowledge carrying information about the correct assignment of a subset of trajectories representing the desired segments.

We integrate the prior knowledge into our segmentation algorithm by manually labeling a small subset of trajectories. For all experimental results in the followings, for each sequence *two* frames (1 and 500) are manually labeled and fed to the *basic moseg*, while there is only *one* manually labeled frame (the first frame) used in *improved moseg*. Frames 250 and 750 are considered for evaluation. 5% of the trajectories are being priory labeled for the *basic moseg*, and just 2.6%, for *improved moseg*. Considering the segmentation difficulty of this application, these numbers are quite small.

Because [9] is the only semi-supervised motion segmentation approach, for the sake of a more profound comparison, we also created a naive baseline. To do so, we use the same prior knowledge as the *basic moseg* and without applying any segmentation method the results in 250th and 750th frames are compared with the ground truth. To obtain a measure of segmentation accuracy we calculate the intersection over union (IOU). This measure is shown for four cases in Fig. 4; *basic moseg* with one and two priory labeled frames, the baseline, and the *improved moseg* with just one priory labeled frame. Poor results of the baseline indicates the level of segmentation complexity. The *basic moseg* with one labeled frame as a priory has gained 77.62% segmentation accuracy, increasing the priory knowledge to two labeled frames boosts up the results to 94.92%. The *improved moseg* outperforms the basic one by obtaining 96.34% accuracy although it just uses only one labeled frame as prior knowledge.

For a more profound study, segmentation results of *improved moseg* for one of the sequences are shown

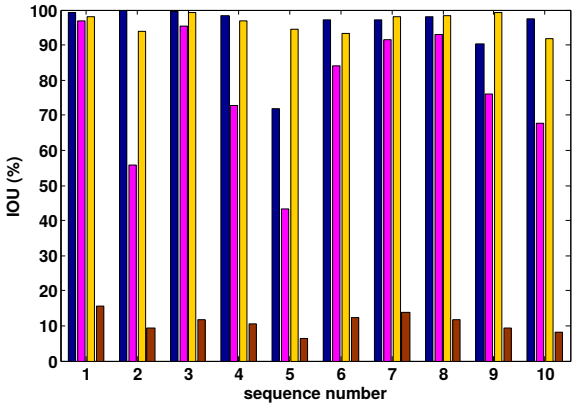


Fig. 4: IOU for different sequences. Given are the results for the *basic moseg* with one manually labeled frame in magenta, two manually labeled frames in blue, *improved moseg* with one labeled frame in yellow and for the baseline in brown.

in Fig. 5 alongside with results of the baseline. It is clear that the baseline in itself performs poorly, while our method exhibits reliable performance.

Occlusion is a longstanding problem in motion segmentation. Frame 650 of Fig. 5 shows a case of severe occlusion where the head is occluded by both hands. As can be seen, the segmentation remains correct and in frame 950, the new trajectories in the occluded area on the head are labeled correctly. Partial occlusion is less of a problem for our proposed method: there are some trajectories left that can still stand in for the terminated ones. These are joined by novel trajectories upon the reappearance of the previously occluded region. In case of a complete occlusion, trajectories could be linked to each other by using rematching process in *improved moseg* that extends the initial labels.

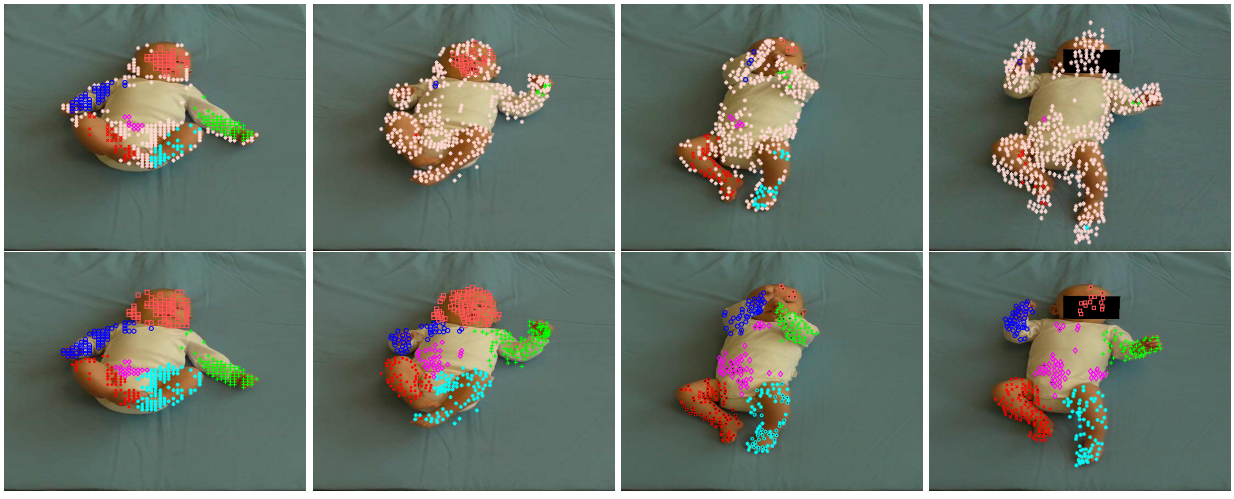


Fig. 5: Seq. 1 segmentation results for frames 1, 200, 650 and 950 from left to right. top row shows the results for the baseline and bottom row is the results for the improved moseg method.

In sequence 5 the infant rolls to the right side, as a consequent many of the priory labeled trajectories terminated and this weakens the *basic moseg* performance, however the *improved moseg* overcomes this problem by re-detecting matched particle when the infant goes to a normal situation. This could be seen in

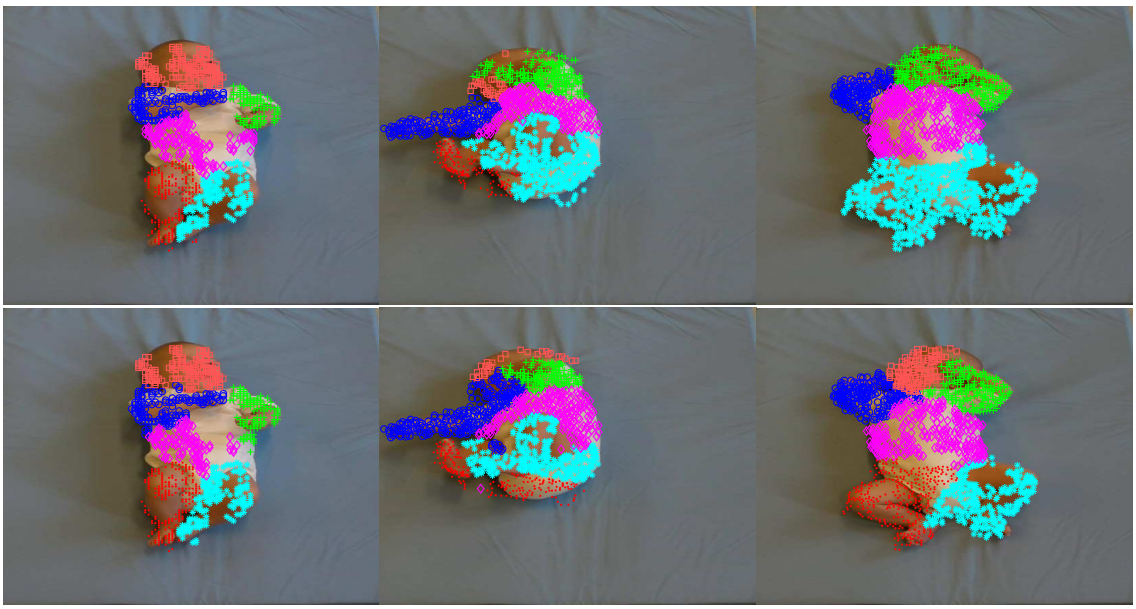


Fig. 6: Segmentation results of sequence 5 in frames 1, 550 and 750. The upper row is for the *basic mosec* with frames 1 and 500 as prior knowledge, and the lower row for *improved moseg* with only frame 1 as prior knowledge.

Fig. 6 where at frame 550 during rolling to the right side, the head and the right foot are occluded and the initially labeled trajectories for these segments are lost. Later on at frame 750 where the baby roles back to the normal situation, we could see that the *basic moseg* wrongly assigns the trajectories on the head and the left foot to the left hand and the left foot, respectively. However, the *improved moseg* performs a correct segmentation by re-matching the particles.

4.2 Comparison with Standard Benchmarks

In this section we challenge our segmentation method with different subjects in order to investigate its generality. The three video sequences *cats02*, *cats04* and *ducks01* of the *Freiburg-Berkeley* data set [1, 6] are considered, in which we deal with occlusion, disocclusion, camera motion, fast motion and low texture objects. As we consider the benchmark of [6], we also employ their metric. Here, we use F-measure that is a region-based metric combining accuracy and coverage of the ground truth [6]. The F-measure between each segmented region c_i and each ground truth region g_i is defined as:

$$F_{i,j} := \frac{2|c_i \cap g_j|}{|c_i| + |g_j|}. \tag{5}$$

Where $|.|$ denotes the size of each set. We define the final metric to be the average F-measure. Fig. 7 shows this measure as a quantitative comparison between the performance of our method, *basic moseg*, the baseline, and [1]. The baseline and [1] show a poor performance. On the other hand, both of basic and *improved moseg* boost the segmentation performance significantly. It should be noticed that in *cats04* one of the cats has a low texture and almost no trajectories is initialized on it, and this aggravates the quality of our methods.

5 Conclusions

In this paper we used motion information as well as prior knowledge of the objects of interest to segment them. The prior knowledge is provided by manually labeling the initial frame, this assigns a sample of

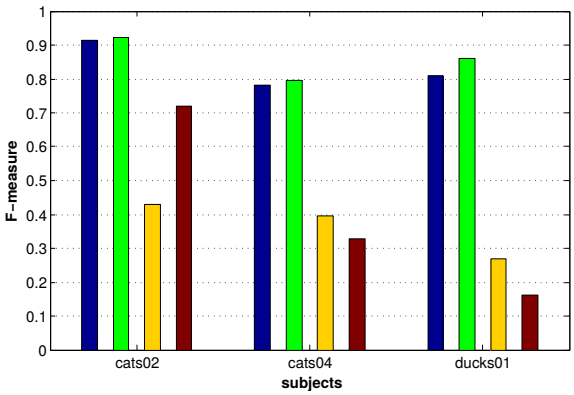


Fig. 7: The F-measure for *basic moseg* in blue, the *improved moseg* in green, [1] in yellow, and the baseline in brown. The middle frame of each sequence is manually labeled and used as prior knowledge.

trajectories to each of the segments. In order to decrease the dependency of the segmentation method on the amount of prior knowledge, a particle matching technique is proposed that uses multi-scale optical flow to re-detect particles through time. Quantitative and qualitative experimental results showed not only the reliability of our method in segmenting different objects, but less dependency on the amount of prior knowledge.

References

Brox, T., Malik, J. (2010). Object segmentation by long term analysis of point trajectories. In ECCV, 282–295.

Cheriyadat, A.M., Radke, R.J. (2009). Non-negative matrix factorization of partial track data for motion segmentation. In ICCV, 865–872.

Dragon, R., Rosenhahn, B., Ostermann, J. (2012). Multi-scale clustering of frame-to-frame correspondences for motion segmentation. In ECCV.

Lezama, J., Alahari, K., Sivic, J., Laptev, I. (2011) Track to the future: Spatio-temporal video segmentation with long-range motion cues. In CVPR, 3369–3376.

Li, Z., Guo, J., Cheong, L.F., Zhou, S.Z. (2013). Perspective motion segmentation via collaborative clustering. In ICCV.

Ochs, P., Malik, J., Brox, T. (2013). Segmentation of moving objects by long term video analysis. TPAMI.

Proesmans, M., Van Gool, L., Pauwels, E., Oosterlinck, A. (1994). Determination of optical flow and its discontinuities using non-linear diffusion. In Computer VisionECCV'94 (pp. 294–304). Springer.

Rahmati, H., Aamo, O.M., Stavdahl, Ø., Dragon, R., Adde, L. (2014). Video-based early cerebral palsy prediction using motion segmentation. In Engineering in Medicine and Biology Society (EMBC), 2014 36th Annual International Conference of the IEEE.

Rahmati, H., Dragon, R., Aamo, O.M., Van Gool, L., Adde, L. (2014). Motion segmentation with weak labeling priors. In GCPR.

Shi, F., Zhou, Z., Xiao, J., Wu, W. (2013). Robust trajectory clustering for motion segmentation. In ICCV.

Sundaram, N., Brox, T., Keutzer, K. (2010). Dense point trajectories by gpu-accelerated large displacement optical flow. In Computer Vision–ECCV 2010 (pp. 438–451). Springer.

Tron, R., Vidal, R. (2007) A benchmark for the comparison of 3D motion segmentation algorithms. In CVPR.

Vidal, R., Hartley, R. (2004). Motion segmentation with missing data using powerfactorization and GPCA. In CVPR, pp. 310–316.