Motion Segmentation by Semi-Supervised Classification in Dynamic Scenery

Petr Pulc[1], Oliver Kerul-Kmec[2], Tomáš Šabata[2], and Martin Holeňa[1]

[1] Institute of Computer Science of the Czech Academy of Sciences, Prague, Czech republic
(pulc,martin)@cs.cas.cz
[2] Faculty of Information Technology, Czech Technical University in Prague, Prague, Czech Republic
(keruloli,sabattom)@fit.cvut.cz

Abstract. Automatic description of multimedia content heavily relies on the ability to discover a structure in such data. As our current focus is given to efficient multimedia indexing, we are mainly interested in discovery and segmentation of foreground objects from the background and their respective description or classification.
Although many approaches based on Convolution Neural Networks have emerged lately, they are usually executed on all frames from the media separately which is, to our belief, wasteful and poorly scalable. On the other hand, methods based on visual Simultaneous Localisation and Mapping (visual SLAM) utilise the temporal structure of the motion picture to extract at first a model of the environment and objects in the scene and later pass these models to methods for object description.
In this paper, we will discuss the first two parts of the visual SLAM – motion tracking and segmentation. While many approaches impose strict restrictions in the segmentation phase to filter motion tracking outliers, we introduce restrictions to the motion tracking itself. Such approach enables us to use off-the-shelf semi-supervised classification methods in the motion segmentation phase without explicit outlier filtering.

Keywords: feature detection, motion tracking, motion segmentation

1 Introduction

With the ever-increasing amount of multimedia content, higher resolution and framerate, the requirement for faster multimedia description approaches is more significant than ever. Although the CPU power is more readily available and the image processing tools utilise GPUs much better, these advances are commonly used in favour of devising more complex processing methods rather than improving the existing approaches.

The Simultaneous Localisation and Mapping (SLAM) approach, originating from the computer vision and robotics research [15], is on the other hand aimed at rapid scene reconstruction that allows using high-level scene and object information for successive description. Also, such object description does not have

to be carried out on each frame, opposed to the approach of convolutional neural networks, such as [26]. In an ideal case, the reconstructed objects can be described once and simply tracked throughout the following frames.

To this end, SLAM proposes a pipeline consisting of a motion tracking, motion segmentation and 3D reconstruction to obtain the visual representation of individual objects from consecutive frames and map them into a 3D space. In this paper, we will focus on the first two phases of the SLAM pipeline, as 2D reconstruction is currently sufficient for our goal, whereas 3D reconstruction of static objects [13, 19, 20, 21, 22, 33] and 3D tracking of dynamic objects [2, 11, 24, 25, 34] is extensively studied.

The major challenge is that we deploy visual SLAM in dynamic environments. Therefore, in the motion segmentation step, we need to account for the fact, that not only is the camera moving through the environment, but also that the individual objects are moving independently. As a result, nor the approach based on 8-point correspondency under epipolar geometry discussed in [16] nor the 5-point relative pose resolver [23] can be used directly. They may be utilised to derive the motion of camera (ego-motion) once we know that the segmented objects are static, but we have not segmented the objects yet.

Moreover, the objects may not be rigid. In such cases, the movement of individual objects cannot be modelled by a simple homography from one frame to another. On the other hand, interest point matching based solely on the similarity of their description tends to produce many outliers that would need to be filtered.

In our approach, we propose a method that circumvents the outliers directly in the interest point matching. To this end, we still depend heavily on the similarity of interest point descriptions, but we propose a limitation on a position of the point of interest in the incoming frame. The proposed point matching is, therefore, expected to be faster as well. Consequently, this allows much higher flexibility in selection of the motion segmenter. In our approach, we picked a semi-supervised classifier with cluster regularisation [31].

In the next section, we discuss motion tracking algorithms and propose a significant improvement of our motion tracking algorithm in Section 3. Semi-supervised classifier for motion segmentation is discussed in Section 4. Finally, preliminary results of the motion segmentation based on our motion tracker are presented in Section 5.

2 Motion Tracking

As RGB video content provides no additional information than colour intensities of the individual pixels, motion tracking is estimated by tracking of small pixel patches from one frame to another. The comparison is carried out on a single channel representing luminance of the pixel (as defined in [10]) to simplify the search for a corresponding patch in the consecutive frame. If the similarity of the patches is based directly on the luminance values from the pictures, the detected apparent motion of these patches is known as optical flow.

The optical flow can be then detected on all pixels and their corresponding patches. This results in a dense optical flow [8], which may be unstable on larger areas without visible structure and is rather expensive to compute.

To increase the stability of optical flow approaches, Shi and Tomasi proposed in [30] a method of selecting only the points from the first considered frame that are supposed to be easily traceable – for example surrounded by significant luminance changes. In combination with an iterative image registration technique proposed by Lucas and Kanade [18] that utilises an image pyramid and an iterative gradient search method, correspondences of individual proposed points of interest are gathered even if the spatial distance of the new patch is greater than one pixel.

This combination of feature detector and optical flow detector is commonly referred to as a Kanade-Lucas-Tomasi feature tracker and present a viable alternative for image sequences with subtle motion and, therefore, short motion vectors as a result. The key issue with this tracker is, however, that the algorithm tries to get correspondences for all points of interest, although these points may disappear from the frame or be occluded. Such misdetections need to be filtered by executing the feature tracker backwards and checking the stability of the proposed motion vector, which slows down the whole process.

In the report by Torr and Zisserman [32], an alternative approach to feature tracking is discussed. Instead of discovering matches based on correspondences of brightness values, they advocate the use of a robust feature descriptor to capture the local structure around the point of interest in both frames. Matching the points of interest from one picture to another is then based solely on the similarity of the interest point descriptions.

Scale-Invariant Feature Transform (SIFT) [17] and Speeded-Up Robust Features (SURF) [3] are the two most widely used feature descriptors that include their own interest point detectors. However, even the improved algorithm is still computationally expensive. For time-critical applications, points of interest detected by Features from Accelerated Segment Test (FAST) [27] combined with a Binary Robust Independent Elementary Features (BRIEF) [5] or Binary Robust Invariant Scalable Keypoints (BRISK) [14] is considered more appropriate. One such combination is the ORB: Oriented FAST and Rotated BRIEF [28] which presents a decent alternative to SIFT and SURF with significantly lower computational requirements. As the GPU accelerated implementation of ORB is already available in the OpenCV library [4], it represents a perfect feature extractor for use in real-time applications and also for our research.

To find out corresponding pairs of extracted features, a distance between the feature descriptions has to be found. SIFT and SURF descriptors are real-valued and, therefore, Euclidean or Mahalanobis distance is used. Binary descriptors (BRIEF and BRISK) simplify the distance measurement to a Hamming distance that can be computed directly with `xor` and `popcnt` CPU instructions.

The most serious issue with the matching based solely on the distance of feature descriptor is that similar points of interest on unrelated positions may be mistaken with each other, which results in false correspondences. Visual SLAM in

static environments, therefore, uses robust position estimators based on sample consensus (such as Random Sample Consensus, RANSAC [9]) to reject matches not following the picture homography.

Unfortunately, this approach is not viable in dynamic environments, as the objects moving in the foreground would also violate the homography estimated from the background and all matches on the foreground objects will be thus eliminated as outliers.

3 Hierarchical Tracking of Smooth Motion

Taking into account the limitations mentioned in the previous section and the requirement to process 4K video in real-time, our approach to the extraction of points of interest and their matching is based on the following concepts:

- Speed is crucial. Therefore GPU accelerated ORB is used for interest point detection and description. Other operations can be executed on a CPU.
- Matching of the interest points needs to be based on both the point description and the estimation of new feature position in the next frame.

To this end, our current algorithm of real-time interest point matching, making an assumption of smooth object motion, consists of following steps:

1. ORB features are computed on three level scale pyramid for the new frame.
2. A fourth virtual level containing a global homography is constructed using the RANSAC from a top-most layer on the current and the new frame.
3. Matching of interest points from the current frame (filled circles in Fig. 1) to the new one (dashed circles) is carried out for the three pyramid layers from the coarsest to the finest:
   a) For each feature from the current layer (point **a**) an estimation of new position (a_n) is computed with respect to the motion of the nearest point in the layer above (A_n) and relative previous motion of the feature ($a_{n-1} - A_{n-1}$). Thus, $a_n = A_n + a_{n-1} - A_{n-1}$. If the information on previous motion is not available, $a_n = A_n$.
   b) Matching based on the Hamming distance of the feature descriptors (represented by colour of circles) is then limited only to neighbourhood of the estimated position (grey dashed circle). Considering the situation from Fig. 1, point **d** would be selected as the best match, although point **e** has smaller Hamming distance (more similar colour) and point **c** is closer to expected position.

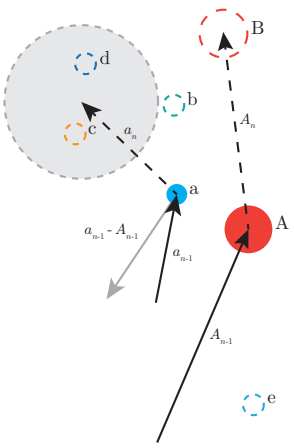

Fig. 1: Feature matching

This algorithm allows us to track motion in complex scenes with many moving objects without a need to derive the optical flow and with minimal risk of severe errors in the motion tracking.

4 Motion Segmentation based on Semi-supervised Learning

Traditionally used methods for motion segmentation in SLAM classify the gathered points of interest into two groups – static and dynamic – to base the approximation of ego-motion only on the static features. Although we need to allow more clusters in the data to segment individual objects, the basic principles of the motion segmentation are the same: to discover a mapping of individual features to the objects in considered scene and background, respectively.

Considering the state-of-the-art methods that are based on the features extracted from optical flow, only the method by Klappstein [12] can segment monocular camera recording. This method uses a graph-cut algorithm on gathered motion metric and is, therefore, computationally intensive (as outlined in [29]) but able to account for some uncertainties in the detected optical flow. Other methods (such as [1, 6, 7]) require a signal from the stereo camera to partially reconstruct depth and are, therefore, unsuitable for our purpose.

Our method takes the same underlying assumption as [12] that the individual objects can be segmented by the position and apparent motion of the points of interest. However, as the motion vectors detected by our approach already account for the significant uncertainties in the optical flow, more straightforward methods of feature space segmentation can be used. Also, our feature space has a low dimensionality (only the x and y feature coordinates, length of the motion vector and its direction), which allows us to use off-the-shelf methods directly.

When a history of motion is not available for any point of interest, unsupervised hierarchical clustering has to be used to propose segments representing individual objects in the scene. However, with the approach discussed in the previous section, such situations should not occur elsewhere than at the processing of the first two frames. During the processing of remaining frames, points of interest retain their label from frame to frame. Therefore, during the analysis of the following frame of the media, the labelled data from the previous one are used for training a semi-supervised classifier.

For motion segmentation, we propose to use the Semi-supervised Classifier with Cluster Regularisation [31] that provides a good generalisation even for small sets of labelled training data.

In our setup, we use a k-means clustering algorithm to provide an initial clustering. Based on these clusters, a pairwise penalty based on correlation is computed to disallow the boundary of final labels in high-density regions. Next, the unlabelled interest points are assigned an initial pseudo-label based on the available labels in the clusters. The computed pairwise penalty is also used to find the ten nearest neighbours of each point of interest.

For the classification, a neural network with one hidden layer is trained. The hidden layer consists of as many neurons as is the number of initial clusters, and output neurons correspond to the considered final labels. The output activation function is softmax, and the minimised loss function is cross-entropy.

5 Experimental Evaluation & Future Work

For our experimental evaluation, we prepared a set of RAW 4K videos (previews available at `https://archive.org/details/motion_ds`) that contain one or two objects with distinct colour on a bright background. This allowed us to extract the ground truth of pixel-to-object correspondence easily. For evaluation, the video was compressed using the H.264 codec to simulate a real-world scenario.

As our main goal is to validate the approach of motion segmentation based on our interest point matching approach and a semi-supervised classification, we devised an experiment concerning the quality of labels proposed by the classifier depending on the delay between classifier training (based on the ground truth data) and testing. To eliminate the possible dependency of the measures on the visual qualities of an individual frame, we repeated the experiment for the first five frames of the video separately.

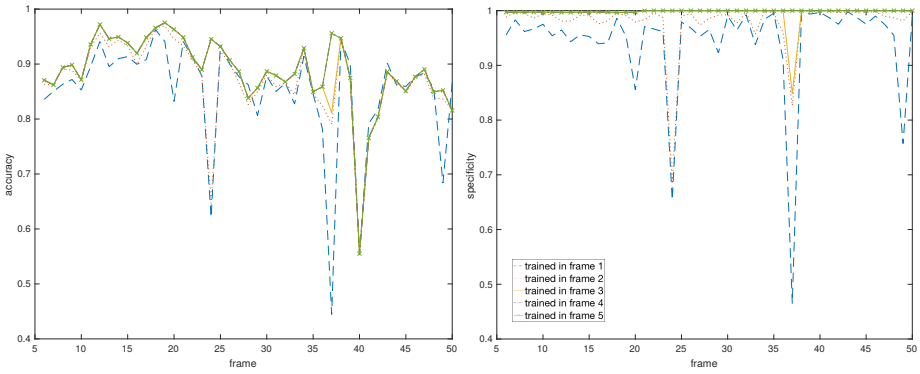

Fig. 2: Accuracy and specificity of the semi-supervised classifier with respect to the selection of training frame. File available at: `https://archive.org/details/motion_ds/handheld_blurbg.mp4`.

Figure 2 indicates that classifiers trained on later frames tend to have higher accuracy and specificity, but in general, the differences between classifiers trained in different frames are insignificant, which we have confirmed by the Friedman test. This behaviour is similar in all considered multimedia files and may be attributed to the fact, that the used codec introduces significant compression artefacts on the first frame (that uses only intraframe compression with limited bandwidth). Later frames have more information available and thus seem to be more stable for interest point detection.

Although motion detection based on optical flow in multimedia and motion segmentation are not novel, in this paper, we have presented a new approach in both areas that ultimately lead to the ability of real-time motion segmentation on 4K video with promising results. While the detection and tracking of 500 points on a Xeon E3-1230 CPU with the Kanade-Lucas-Tomasi algorithm takes

242 ms per frame, our approach on CPU only takes 95 ms. When nVidia 1050 Ti GPU is utilised, processing of a single 4K frame takes only 35 ms on average, which is just enough for real-time processing of 25 fps media.

In future, we intend to transfer all data-intensieve operations to a GPU and more importantly, to validate our approach on nonsynthetic datasets.

Acknowledgements

The work has been supported by the grant 18-18080S of the Czech Science Foundation (GACR).

References

[1] Pablo F Alcantarilla et al. "On combining visual SLAM and dense scene flow to increase the robustness of localization and mapping in dynamic environments". In: *ICRA*. IEEE. 2012, pp. 1290–1297.

[2] Shai Avidan and Amnon Shashua. "Trajectory triangulation: 3D reconstruction of moving points from a monocular image sequence". In: *TPAMI* 4 (2000).

[3] Herbert Bay et al. "Speeded-up robust features (SURF)". In: *Computer vision and image understanding* 110.3 (2008), pp. 346–359.

[4] G. Bradski. "The OpenCV Library". In: *Dr. Dobb's Journal of Software Tools* (2000).

[5] Michael Calonder et al. "Brief: Binary robust independent elementary features". In: *ECCV*. Springer. 2010, pp. 778–792.

[6] Maxime Derome et al. "Moving object detection in real-time using stereo from a mobile platform". In: *Unmanned Systems* 3.04 (2015), pp. 253–266.

[7] Maxime Derome et al. "Real-time mobile object detection using stereo". In: *ICARCV*. IEEE. 2014, pp. 1021–1026.

[8] Gunnar Farnebäck. "Two-frame motion estimation based on polynomial expansion". In: *SCIA*. Springer. 2003, pp. 363–370.

[9] Martin A Fischler and Robert C Bolles. "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography". In: *CACM* 24.6 (1981), pp. 381–395.

[10] ITU-R. *Studio encoding parameters of digital television for standard 4:3 and wide-screen 16:9 aspect ratios*. Recommendation. Geneva, Mar. 2011.

[11] Jeremy Yirmeyahu Kaminski and Mina Teicher. "General trajectory triangulation". In: *ECCV*. Springer. 2002, pp. 823–836.

[12] Jens Klappstein et al. "Moving object segmentation using optical flow and depth information". In: *Pacific-Rim Symposium on Image and Video Technology*. Springer. 2009, pp. 611–623.

[13] Georg Klein and David Murray. "Parallel tracking and mapping on a camera phone". In: *ISMAR*. IEEE. 2009, pp. 83–86.

[14] Stefan Leutenegger, Margarita Chli, and Roland Y Siegwart. "BRISK: Binary robust invariant scalable keypoints". In: *ICCV*. IEEE. 2011.

[15] H. Lim, J. Lim, and H. J. Kim. "Real-time 6-DOF monocular visual SLAM in a large-scale environment". In: *ICRA*. 2014, pp. 1532–1539. DOI: 10.1109/ICRA.2014.6907055.

[16] H Christopher Longuet-Higgins. "A computer algorithm for reconstructing a scene from two projections". In: *Nature* 293.5828 (1981), p. 133.

[17] David G Lowe. "Distinctive image features from scale-invariant keypoints". In: *International journal of computer vision* 60.2 (2004), pp. 91–110.

[18] Bruce D Lucas, Takeo Kanade, et al. "An iterative image registration technique with an application to stereo vision". In: (1981).

[19] Etienne Mouragnon et al. "Generic and real-time structure from motion". In: *British Machine Vision Conference 2007 (BMVC 2007)*. 2007.

[20] Etienne Mouragnon et al. "Generic and real-time structure from motion using local bundle adjustment". In: *Image and Vision Computing* 27.8 (2009).

[21] Etienne Mouragnon et al. "Real time localization and 3d reconstruction". In: *CVPR*. Vol. 1. IEEE. 2006, pp. 363–370.

[22] Raul Mur-Artal, Jose Maria Martinez Montiel, and Juan D Tardos. "ORB-SLAM: a versatile and accurate monocular SLAM system". In: *IEEE Transactions on Robotics* 31.5 (2015), pp. 1147–1163.

[23] David Nistér. "An efficient solution to the five-point relative pose problem". In: *TPAMI* 26.6 (2004), pp. 756–770.

[24] Kemal E Ozden, Konrad Schindler, and Luc Van Gool. "Multibody structure-from-motion in practice". In: *TPAMI* 32.6 (2010), pp. 1134–1141.

[25] Hyun Soo Park et al. "3D reconstruction of a moving point from a series of 2D projections". In: *ECCV*. Springer. 2010, pp. 158–171.

[26] Joseph Redmon and Ali Farhadi. "YOLOv3: An Incremental Improvement". In: *arXiv* (2018).

[27] Edward Rosten and Tom Drummond. "Machine learning for high-speed corner detection". In: *ECCV*. Springer. 2006, pp. 430–443.

[28] Ethan Rublee et al. "ORB: An efficient alternative to SIFT or SURF". In: *ICCV*. IEEE. 2011, pp. 2564–2571.

[29] Muhamad Risqi U Saputra, Andrew Markham, and Niki Trigoni. "Visual SLAM and Structure from Motion in Dynamic Environments: A Survey". In: *ACM Computing Surveys (CSUR)* 51.2 (2018), p. 37.

[30] Jianbo Shi and Carlo Tomasi. *Good features to track*. Tech. rep. Cornell University, 1993.

[31] Rodrigo GF Soares, Huanhuan Chen, and Xin Yao. "Semisupervised classification with cluster regularization". In: *IEEE transactions on neural networks and learning systems* 23.11 (2012), pp. 1779–1792.

[32] Philip HS Torr and Andrew Zisserman. "Feature based methods for structure and motion estimation". In: *IWVA*. Springer. 1999, pp. 278–294.

[33] Changchang Wu. "Towards linear-time incremental structure from motion". In: *3DV*. IEEE. 2013, pp. 127–134.

[34] Enliang Zheng et al. "Joint object class sequencing and trajectory triangulation (jost)". In: *ECCV*. Springer. 2014, pp. 599–614.