Region Covariance Matrix-Based Object Tracking with Occlusions Handling

Ivar Austvoll[1] and Bogdan Kwolek[2]

[1] University of Stavanger, N-4036 Stavanger, Norway
ivar.austvoll@uis.no
[2] Rzeszów University of Technology, 35-959 Rzeszów, Poland
bkwolek@prz.edu.pl

Abstract. This work proposes an optical-flow based feature tracking that is combined with region covariance matrix for dealing with tracking of an object undergoing considerable occlusions. The object is tracked using a set of key-points. The key-points are tracked via a computationally inexpensive optical flow algorithm. If the occlusion of the feature is detected the algorithm calculates the covariance matrix inside a region, which is located at the feature's position just before the occlusion. The region covariance matrix is then used to detect the ending of the feature occlusion. This is achieved via comparing the covariance matrix based similarity measures in some window surrounding the occluded key-point. The outliers that arise in the optical flow at the boundary of the objects are excluded using RANSAC and affine transformation. Experimental results that were obtained on freely available image sequences show the feasibility of our approach to perform tracking of objects undergoing considerable occlusions. The resulting algorithm can cope with occlusions of faces as well as objects of similar colors and shapes.

1 Introduction

Tracking an object in a sequence of images is currently utilized in many computer vision applications. The goal of visual tracking is to locate a region in each image that matches an appearance of a target object. The algorithms for visual object tracking can be divided broadly into two categories, namely: feature-based and visual-model/template-based [1]. Feature-based methods track an object through tracking a collection of local features such as corners [2]. The second group of methods achieves the object tracking through matching a template or a model to the input image [3].

In typical scenarios, interactions between moving objects result in partial or significant occlusions, making the object tracking a highly challenging problem. Various systems and methods have been proposed to handle object tracking in complex crowded scenes with the occlusions arising in the object tracking [4]. Multiple cameras are often used to cope with occlusions [5]. In most conventional multi-camera systems the targets are represented as a collection of blobs in 3D space, which are tracked over time. This requires finding the corresponding blobs

across multiple images as well as assigning 2D blobs to the 3D blobs. However, using a multi-camera system in many applications may be impractical. Therefore, stereo cameras are often used to perform object tracking in such circumstances [6]. However, conventional stereo cameras usually do not provide useful and reliable depth estimates in occluded regions, particularly when they are textureless.

The object tracking is often achieved using a single camera. However, one fundamental limitation of using one camera in the tracking of objects is dealing with object occlusions. In single-camera methods, occlusion can be identified through prediction of the object location or on a per-pixel basis. Kalman filtering or particle filtering [7] can be used to predict the positions of objects during occlusions. Methods relying on per-pixel representation often use templates to represent objects. The underlying main assumption behind template matching is that the appearance of the template remains almost the same throughout the entire image sequence. Hence, handling occlusions is not an easy task in such an approach [3]. Babenko *et. al* [8] recently proposed an online multiple instance learning algorithm to achieve robust object tracking under occlusion. However, to achieve long-term object tracking, a persistent tracker must cope with occlusions as well as must be able to reacquire the object in case of considerable occlusions.

Despite the above advances, in many situations the existing algorithms do not have satisfactory tracking robustness, especially when there is a large amount of occlusion between two or more objects. Therefore, a highly efficient occlusion handling scheme, which could lead to a considerable improvement of the tracking performance, even when there is a large amount of occlusion between two or more objects is needed. Our approach to cope with considerable occlusions is to construct a region covariance matrix in the surround of the feature just before occlusion and then to employ such a descriptor to detect the ending of the occlusion. The motivation behind such an approach is that the region covariance matrix is a strong and robust indicator for point-to-point correspondence. It is a powerful descriptor that encodes the variance of the channels, such as red, green, blue, gradient, etc., their correlations with each other, and the spatial layout [9]. Moreover, variations in illumination as well as in pose or viewpoint do not affect the covariance considerably. Through the use of such a robust region descriptor a tracked feature can be detected and recognized again after loosing it.

The features are tracked using the optical flow. In optical flow-based feature tracking the significant errors might occur on the occluding boundary. In [10] it has been shown that features belonging to the same object have correlated behavior, whereas features belonging to different objects show evidence of more uncorrelated and independent behavior. Motivated by this observation we assume that key-points in previous and current images are related by an affine transformation, and we then try several combinations of 3 points in a RANSAC framework to exclude outliers, i.e. features that do not move consistently with the inliers. This helps us to detect the occlusion, and more importantly, the features do not undergo undesirable shifting through the appearance changes at the boundaries where occlusions take place.

In Section 2 we present a feature-based object tracking and start with a discussion of feature detection and optical flow estimation. Thereafter we present how the consistency of matches is handled in our approach. Then we outline covariance matrix based region descriptor as well as present our algorithm. Section 3 is devoted to demonstration of experimental results. We end the paper with conclusions.

2 Feature-Based Object Tracking

2.1 Feature detection and optical flow estimation

Good detectors of features are very important for object tracking [2]. Several feature detectors and descriptors have been proposed and evaluated in the literature [11][12]. Recent research [12] has demonstrated that the repeatability of the key-point detectors deteriorates with change of the viewpoint. The work mentioned above has also demonstrated that no key-point detector performs well in case of considerable view changes. Taking this into account we utilize the Harris corner detector [13] in our algorithm. Another rationale of our choice is that the Harris corner detector has relatively low computational cost when compared to the SIFT algorithm.

The inter-frame translations of the key-points are determined by the Lucas-Kanade optical flow algorithm [14]. This method is still one of the best methods for two-frame motion estimation. The advantage of the method is that the features can be tracked with low computational cost and therefore it is utilized in our algorithm.

2.2 Consistency of matches

RANSAC (RANdom SAmple Consensus) is a robust method to estimate parameters of a mathematical model from a set of data contaminated by considerable amounts of outliers [15]. The percentage of outliers which can be handled by RANSAC can be larger than 50% of the entire data set. The RANSAC algorithm consists of two steps, which are repeated in an iterative hypothesis-and-test fashion. In the hypothesis step, minimal sample sets are randomly chosen from the input dataset and then the model parameters are estimated using only elements from such sets. In the second step, the RANSAC tests whose elements of the entire dataset are consistent with a model, which is instantiated with the parameters from the first step. The steps mentioned above are repeated a fixed number of times. Each time, either a refined model is produced or the model is declined because too few points are classified as inliers.

In our approach we employ RANSAC to find the largest set of matches consistent with an affine transformation. The affine transformation is given by the following equation:

$$\begin{bmatrix} x_i' \\ y_i' \\ 1 \end{bmatrix} = \begin{bmatrix} h_1 & h_2 & h_3 \\ h_4 & h_5 & h_6 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} x_i \\ y_i \\ 1 \end{bmatrix} \tag{1}$$

where $[x'_i \;\; y'_i \;\; 1]$ is the matched feature location in the current image and $[x_i \;\; y_i \;\; 1]$ is the matched feature location in the previous image. A sum of square errors is minimized to estimate the affine transform parameters $h = [h_1 \;\; h_2 \;\; h_3 \;\; h_4 \;\; h_5 \;\; h_6]$ according to the following equation:

$$\min_h \sum_i (h_1 x_i + h_2 y_i + h_3 x'_i)^2 + (h_4 x_i + h_5 y_i + h_6 y'_i)^2 \tag{2}$$

The solution to the optimization problem (2) is given by the following equation:

$$h = A^{-1} b \tag{3}$$

where

$$A = \sum \begin{bmatrix} x_i^2 & x_i y_i & x_i \\ x_i y_i & y_i^2 & y_i \\ x_i & y_i & 1 \end{bmatrix} \tag{4}$$

and

$$b = \sum [x'_i x_i \;\; x'_i y_i \;\; x'_i \;\; y'_i x_i \;\; y'_i y_i \;\; y'_i]^T \tag{5}$$

2.3 Covariance matrix based region descriptor

Recently, in [16] an elegant and simple solution to integrate multiple image features has been proposed. It is based on the covariance matrix. Using a covariance matrix (CM) as a region descriptor has many advantages, namely: 1) CM indicates both spatial and statistical properties of the objects; 2) it provides an elegant means to combine multiple modalities and features; 3) it is capable of relating regions of different sizes.

Let $\mathbf{I}$ be an image of size $W \times H$. At each pixel location $\mathbf{x} = [x, y]^T$ we can extract d features such as intensity, gradient, color, etc. Using such a feature set we can construct a $W \times H \times d$ feature image $\mathbf{H}$. Given a rectangular window R we can then compute the covariance matrix $\mathbf{c}_R$ of the features according to the following equation:

$$\mathbf{c}_R = \frac{1}{|R| - 1} \sum_{\mathbf{x} \in R} (\mathbf{H} - \mathbf{m}_R)(\mathbf{H} - \mathbf{m}_R)^T \tag{6}$$

where $\mathbf{m}_R = \frac{1}{|R|} \sum_{\mathbf{x} \in R} \mathbf{H}(\mathbf{x})$ denotes the vector of means of corresponding features for the pixels in region R, and $|R|$ stands for the size of region R. The diagonal entries in such a covariance matrix express the variance of each feature and the off-diagonal entries indicate their mutual correlations. The covariance matrix is a very informative region descriptor because it encodes information about the variance of features, their correlations witch each other, and spatial layout. It can be computed efficiently through the use of integral images in a way that has been shown in [9]. To measure the dissimilarity between the covariance matrixes $\mathbf{c}_1$ and $\mathbf{c}_2$ we employed the following distance [16]:

$$\rho(\mathbf{c}_1, \mathbf{c}_2) = \sqrt{\sum_{i=1}^{|R|} \ln^2 \lambda_i(\mathbf{c}_1, \mathbf{c}_2)} \tag{7}$$

where $\{\lambda_i(\mathbf{c}_1, \mathbf{c}_2)\}_{i=1,\ldots,|R|}$ are the generalized eigenvalues of $\mathbf{c}_1$ and $\mathbf{c}_2$, which are calculated on the basis $\lambda_i \mathbf{c}_1 \mathbf{x}_i - \mathbf{c}_2 \mathbf{x}_i$, where $\mathbf{x}_i \neq 0$ are the generalized eigenvectors. Another possibility to measure the similarity between covariance matrices is to use Log-Euclidean metrics [17].

2.4 The algorithm

When a new frame is available, after detecting Harris corners, the optical flow is estimated to determine the current location of key-points. At this stage we calculate the quality of the features [2] in order to verify if they are still trackable and have not drifted away from original targets. Through monitoring the features' quality we verify whether each feature is occluded or not. In case of an occlusion we calculate the region covariance matrix at feature's location before the occlusion and afterwards we finally decide if the occlusion takes place. For the non-occluded features we apply RANSAC with the affine model in order to determine the outliers, i.e. features that move inconsistently according to the best affine model. For such features we compute region covariance matrixes and add the features to the set of occluded features. This way we suppress the motion errors that arise at the boundary of the occlusions. In subsequent frames for each occluded feature we execute a test if the occlusion is finished. Given the feature location before the occlusion we perform greedy search in a window surrounding such a location for the best similarity of region based covariance descriptors. If the best distance between covariance matrixes is below the threshold we start the tracking of the feature. In [18] the RANSAC algorithm is used to identify consistent subsets of correspondences and obtain a better homography. Our work differs from the mentioned work in that we focus on handling the occlusions of the object undergoing tracking with the support of the RANSAC.

3 Experimental Results

We validated the algorithm by tracking a face, which undergoes considerable occlusions*. Although almost the whole face was occluded our tracker successfully tracks the face, see images in upper row of Fig. 1, as well as detects it after the occlusion, see images in bottom row. As we can observe the algorithm is able to reacquire the object despite similar colors as well as textures of both objects. The above mentioned similarity of both objects leads to sporadic misdetections of the feature's occlusion as it can be seen in frame #7, where some features are located at the occluding hand. Through the use of RANSAC built on an affine motion model, such feature drifting is eliminated quickly as can be observed in frame #9. In frame #13 we can notice that the algorithm reacquired most of the features. The location of the features is consistent with their location before

* Thanks Dr. Birchfield for this sequence, obtained from http://robotics.stanford.edu/~birch/headtracker

occlusion, see frame #5. In the next frames of the discussed sequence we can perceive the behavior of the algorithm after redetection of the features and during the subsequent occlusion. As we can see at frame #26 the number of redetected features is sufficient to continue the tracking of the face. The locations of the redetected features are consistent with the location before the second occlusion, see frame #13, as well as with initial feature locations, see frame #5.

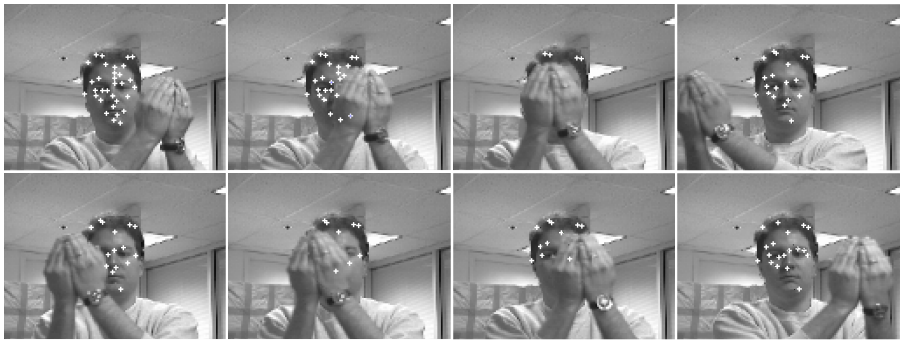

Fig. 1. Object tracking during considerable occlusion. Frames #5, 7, 9, 13 (upper row), #20, 21, 23, 26 (bottom row)

The experimental results shown in Fig. 2 demonstrate that the RANSAC algorithm allows us to obtain far better tracking results. In particular, as experimental results show, without RANSAC some features undergo undesirable shifts and in consequence the number of the reacquired features is somewhat smaller, see frame #26. Even more, as we can see at the mentioned image, some features can be located in the wrong objects.

Fig. 2. Key-point-based object tracking without RANSAC. Frames #7, 9, 23, 26

Figure 3 demonstrates a scene in which a face undergoing tracking is temporally occluded by another one. The occluded face moves slightly during the occlusion and in consequence the number of the reacquired features is something smaller. The extension of the algorithm about Procrustes analysis [19] to compute the similarity transform (translation, rotation and scale) between two sets of visible features and then to transform the occluded features is straightforward.

Fig. 3. Tracking a face that undergoes occlusion by an another face. Frames #420, 422, 434, 449

The algorithm has been implemented in Matlab. The above described experiments were done on color images of size 128×96. The recovery of the occluded feature is done through the greedy search for the best similarity of the region covariance matrixes and then comparing it with a threshold value. The searching is realized in a window of size 5×5 centered on feature's position just before the occlusion. The region covariance is built in a windows of size 6×6 using feature location, R, G, B color values and first and second image derivatives.

4 Conclusions

To persistently track an object in long image sequences the algorithm must cope with considerable occlusions. Since the existing algorithms can not perform well under considerable occlusions, we propose an algorithm that employs region covariance descriptors to reacquire the occluded features. We demonstrated experimentally that such a descriptor is very useful in recovering the features. The RANSAC algorithm helps considerably in detecting the occlusions as well as allows us to exclude outliers arising at the boundary between moving objects.

Acknowledgment

Portions of this work were completed by B. Kwolek in the Faculty of Science and Technology at the University of Stavanger during a research stay, which has been supported within the EEA Financial Mechanism and the Norwegian Financial Mechanism.

References

1. Yilmaz, A., Javed, O., Shah, M.: Object tracking: A survey. ACM Comput. Surv. **38** (2006) 13
2. Shi, J., Tomasi, C.: Good features to track. In: Proc. of CVPR. (1994) 593–600
3. Schreiber, D.: Robust template tracking with drift correction. Pattern Recogn. Lett. **28** (2007) 1483–1491
4. Gabriel, P.F., Verly, J.G., Piater, J.H., Genon, A.: The state of the art in multiple object tracking under occlusion in video sequences. In: Int. Conf. on Advanced Concepts for Intelligent Vision Systems. (2003) 166–173

5. Khan, S., Shah, M.: A multiview approach to tracking people in crowded scenes using a planar homography constraint. In: Proc. of the 10th European Conf. on Computer Vision, Graz, Austria (2006) 133–146
6. Darrell, T., Gordon, G., Harville, M., Woodfill, J.: Integrated person tracking using stereo, color, and pattern detection. Int. J. Comput. Vision **37** (2000) 175–185
7. Isard, M., Blake, A.: Condensation - conditional density propagation for visual tracking. Int. J. of Computer Vision **29** (1998) 5–28
8. Babenko, B., Yang, M.H., Belongie, S.: Visual tracking with online multiple instance learning. In: IEEE Comp. Society Conf. on Computer Vision and Pattern Recognition, Miami, Florida, USA (2009) 983–990
9. Tuzel, O., Porikli, F., Meer, P.: Region covariance: A fast descriptor for detection and classification. In: European Conference on Computer Vision, Graz, Austria, Lecture Notes in Artificial Intelligence, vol. 3952, Springer-Verlag Berlin Heidelberg (2006) 589–600
10. Ramanan, D., Forsyth, D.A.: Using temporal coherence to build models of animals. In: Proc. of the Ninth IEEE Int. Conf. on Computer Vision, Washington, DC, USA, IEEE Computer Society (2003) 338–345
11. Lowe, D.G.: Distinctive image features from scale-invariant keypoints. Int. Journal of Computer Vision **60** (2004) 91–110
12. Mikolajczyk, K., Schmid, C.: A performance evaluation of local descriptors. IEEE Trans. Pattern Anal. Mach. Intell. **27** (2005) 1615–1630
13. Harris, C., Stephens, M.: A combined corner and edge detector. In: Proc. of Fourth Alvey Vision Conference, Manchester, UK (1988) 147–151
14. Lucas, B., Kanade, T.: An iterative image registration technique with an application to stereo vision. In: Proc. Int. Joint Conf. on Artificial Intell. (1981) 674–679
15. Fischler, M.A., Bolles, R.C.: Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cartography. Commun. ACM **24** (1981) 381–395
16. Porikli, F., Tuzel, O., Meer, P.: Covariance tracking using model update based on lie algebra. In: Proc. Int. Conf. on Comp. Vision and Pattern Recognition, vol. 1 (2006) 728–735
17. Arsigny, V., Fillard, P., Pennec, X., Ayache, N.: Fast and simple calculus on tensors in the log-euclidean framework. In: Int. Conf. on Medical Image Computing and Computer Assisted Intervention. (2005) 115–122
18. Okuma, K., Little, J.J., Lowe, D.G.: Automatic rectification of long image sequences. In: Asian Conference on Computer Vision (ACCV), Jeju Island, Korea (2004)
19. Mardia, K., Dryden, I.: Statistical Shape Analysis. Wiley (1998)