Performance Evaluation of Object Representations in Mean Shift Tracking

Peter Hosten, Andreas Steiger, Christian Feldmann, and Christopher Bulla
Institut für Nachrichtentechnik
RWTH Aachen University
Aachen, Germany
Email:{hosten, steiger, feldmann, bulla}@ient.rwth-aachen.de

***Abstract*—Mean shift tracking is a real-time capable object tracking approach that is not restricted to a specific object category. Several target object representations based on a feature distribution within an object region have been proposed for mean shift tracking. Quantitative performance metrics for the evaluation of object representations in mean shift tracking are mainly based on a comparison against ground truth data, which is often not available or requires considerable effort for its creation. In this paper, our main contribution is a novel approach for the quantitative evaluation of object representations in mean shift tracking, that does not rely on any ground truth data. Our approach is based on multiple hypotheses for the object location which initialise the mean shift tracking algorithm. The tracking result is then treated as random process and a quantitative metric is derived from its properties. Finally, the evaluation approach is applied to various object representations and test sequences. The findings demonstrate that the usage of multi-part object representations is beneficial if the representation captures the spatial colour distribution of the object.**

***Keywords*- mean shift tracking; multi-part object representation; tracking evaluation**

I. INTRODUCTION

The expansion of mobile networks and the spread of mobile devices allow for a universal multimedia access (UMA) in heterogeneous environments [1]. This requires an adaptation of the multimedia content in order to meet the current user situation such as the available data rate or the display device capability. However, considering only the technical requirements in the adaptation process does not necessarily ensure an optimal user experience. Current developments therefore aim to focus on the user and try to adapt the multimedia content with respect to the user preferences as well. The vision of user-centric convergence of multimedia is generally known as universal multimedia experience (UME) [2].

In this context, video adaptation and presentation techniques have become popular that are guided by region of interest (ROI) information. ROI-based video transcoding, for example, allows to reduce the quality of the different regions according to their importance, whereas ROI-aware rich media presentations allow for the interaction with the ROIs [3].

Consequently, these adaptation systems demand automatically created video annotations. Though a clear definition of an ROI cannot be given in general, it is commonly assumed that video objects might be of interest to the user. An automatic detection of arbitrary objects, however, is infeasible in practise. In principle, objects can only be detected when the underlying model assumptions are met. Thus, object detectors that have been trained for a specific appearance of an object, typically have a limited generalisation ability, e.g. they are not able to handle arbitrary deformations or occlusions. Hence, a reliable detection is generally not possible for the complete video, but for certain frames. In order to fill this gap, tracking approaches are necessary that allow to track the detected object and ROI, respectively.

Object tracking comprises an estimation of the target object state based on previous state estimations and the processing of visual information of the current frame. Though several real-time capable tracking methods have been proposed in literature [4], mean shift tracking is of particular interest as it allows for a generic modelling of the object's appearance by a probability density function (PDF) of features [5] and is thus not restricted to a specific object category. It seeks a mode of a similarity function between the target model and a candidate model by iterative computations of mean shift updates. A widespread feature is colour information whose distribution is encoded by a histogram. In order to gain a more distinct object representation, enhanced object representations for mean shift tracking have been proposed in form of multi-part object regions [6] [7].

In order to investigate the suitability of these object representations, in this work a novel quantitative evaluation method is proposed. Common approaches for the evaluation of tracking algorithms and object representations are based on ground truth data such as object centroids or bounding boxes [8] [9]. Object representations for mean shift tracking have been particularly evaluated based on the dice coefficient and the distance of the tracker centroid to the ground truth centroid [10]. The object centroid does, however, not correspond to the mode of the similarity function which is seeked by the mean shift tracking. Furthermore, the mode of the similarity function varies dependent on the underlying object representation.

ISBN: 978-1-61208-265-3

Therefore, we propose an evaluation approach which is independent on ground truth data and focused on the convergence behaviour of the mean shift tracking for different object representations. Based on multiple tracking initialisations drawn from an input random process, the mode to which mean shift tracking converges is treated as random process. Its stochastic properties are used to derive a metric allowing for an analysis of tracking accuracy and robustness of different object representations.

The rest of this paper is organised as follows: In Section II, the mean shift tracking of the target object location is explained and some object representations are presented. In Section III, we present a novel approach for the performance evaluation of object representations in mean shift tracking. Results are provided in Section IV. Finally, Section V concludes and discusses future work.

II. MEAN SHIFT TRACKING

A. Object Representation

The target object is represented by a target model which comprises the PDF of features within an object region. A target candidate at a candidate location is computed according to the same object representation and is evaluated against the reference target model during the course of tracking. Mean shift tracking based on colour features encodes the PDF of colours by a normalised kernel-weighted M-bin histogram [5] at which the weighting kernel $K(\mathbf{x})$ is centred at the target object. The target model $\mathbf{q} = \{q_u\}_{u=1,\ldots,M}$ and a candidate model $\mathbf{p}(\mathbf{y}) = \{p_u(\mathbf{y})\}_{u=1,\ldots,M}$ at location $\mathbf{y}$ are then computed by:

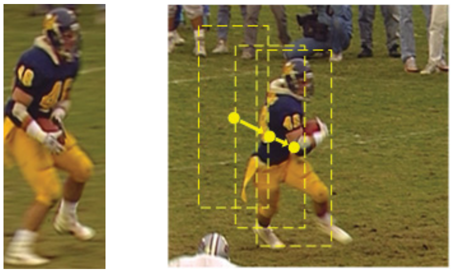
$$q_u = C \cdot \sum_{n=1}^{N} K(\mathbf{y}_o - \mathbf{x}_n)\delta(b(\mathbf{x}_n) - u) \tag{1}$$

$$p_u(\mathbf{y}) = C_h \cdot \sum_{n=1}^{N} K\left(\frac{\mathbf{y} - \mathbf{x}_n}{h}\right)\delta(b(\mathbf{x}_n) - u) \tag{2}$$

Here, u denotes an index of a histogram bin, $b(\cdot)$ yields the bin index of the colour at pixel position $\mathbf{x}_n$, $\delta(\cdot)$ is the Kronecker delta function, N the number of pixels within the object region, and C and C_h are normalisation constants. Since the scale of the object may vary, the width h of the kernel function must be adapted to the size of the object region.

B. Mean Shift Update

Mean shift has been proposed as technique for seeking the mode of a density estimation [11] based on sample observations $\{\mathbf{x}_n\}$ which may be weighted by weights $\{w_n\}$. In the context of video object tracking, the samples $\{\mathbf{x}_n\}$ represent the pixel positions within the object region of the target and the target candidate, respectively.



(a) target (b) target candidates

Figure 1. The new location of the target is estimated by iterative mean shift updates until a maximum number of iterations or convergence are reached.

Based on a kernel function $G(\mathbf{x})$ centred at a location $\mathbf{y}_{j-1}$, a mode estimation $\mathbf{y}_j$ of the weighted kernel density estimation $\hat{f}_{K,h}(\mathbf{x})$ in (3) is provided by the weighted mean shift update in (4).

$$\hat{f}_{K,h}(\mathbf{x}) = \frac{1}{N}\sum_{n=1}^{N} w_n K\left(\frac{\mathbf{x} - \mathbf{x}_n}{h}\right) \tag{3}$$

$$\mathbf{y}_j = \frac{\sum_{n=1}^{N} w_n \mathbf{x}_n G\left(\frac{\mathbf{y}_{j-1} - \mathbf{x}_n}{h}\right)}{\sum_{n=1}^{N} w_n G\left(\frac{\mathbf{y}_{j-1} - \mathbf{x}_n}{h}\right)} \tag{4}$$

The kernel functions $K(\mathbf{x})$ and $G(\mathbf{x})$ are related through their defining profiles $k(x)$ and $g(x)$ at which $g(x) = -k'(x)$ holds and $K(\mathbf{x})$ denotes the shadow of $G(\mathbf{x})$ [11]. The sequence $\{\mathbf{y}_j\}_{j=1,2,\ldots}$ converges to the true mode of $\hat{f}_{K,h}(\mathbf{x})$ [12] which indicates the most likely location of the target object. As depicted in figure 1, mean shift location tracking therefore comprises iterative mean shift updates until a maximum number of iterations or convergence are reached.

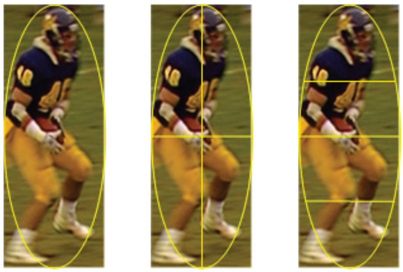
The actual weight w_n at pixel position $\mathbf{x}_n$ is derived from a Taylor series expansion of the Bhattacharyya coefficient $\rho(\mathbf{q}, \mathbf{p}(\mathbf{y}))$ similarity measure between the target model and a candidate model around a candidate model $\mathbf{p}(\mathbf{y}_0)$ [5]:

$$\rho(\mathbf{q}, \mathbf{p}(\mathbf{y})) = \sum_{u=1}^{M} \sqrt{p_u(\mathbf{y}) q_u} \tag{5}$$

$$w_n = \sum_{u=1}^{M} \sqrt{\frac{q_u}{p_u(\mathbf{y}_0)}}\,\delta(b(\mathbf{x}_n) - u) \tag{6}$$

Different approaches for the setting of weights $\{w_n\}$ are however possible such as target model back-projection [13] or various schemes for background incorporation [14].

The mean shift tracking algorithm can be extended to estimate the scale σ and orientation φ by mapping of Cartesian location coordinates $\mathbf{x}$ to a 4-dimensional state space $\Gamma = (\mathbf{x}^\top, \sigma, \varphi)^\top$ and computing the mean shift updates in the 4-dimensional state space [15].



(a) HOL (b) CMP (c) SMP

Figure 2. Illustration of the holistic (HOL), cross multi-part (CMP) and stack multi-part (SMP) object representations.

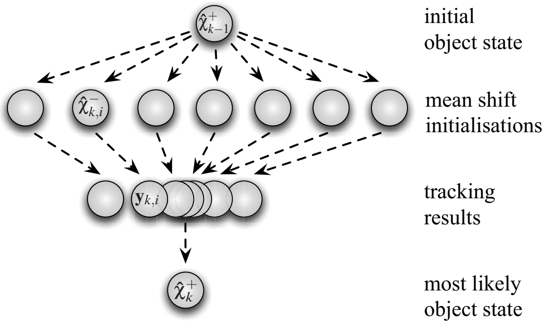


Figure 3. Illustration of Monte Carlo simulation.

C. *Multi-Part Object Representation*

Multi-part object representations divide the object region into subregions to provide a more distinct description of the features in the spatial domain. In contrast to holistic (HOL) object representations illustrated in Fig. 2(a) they provide information about the distribution of features for each subregion of the object region. Various fixed approaches for the spatial division of the object region exist [10]. For our evaluation we consider the cross (CMP) and stack (SMP) approach illustrated in Fig. 2(b) and Fig. 2(c).

III. QUANTITATIVE TRACKING EVALUATION METRIC

The mean shift procedure constitutes a gradient-based mode estimation technique and is therefore only able to locate a local mode of a density function estimation. It is particularly sensitive to different mean shift initialisations which may result in convergence to different modes. Criteria of interest for the evaluation of object models for mean shift tracking include the accuracy of convergence and the robustness of convergence for different initialisations.

In this context, robustness of convergence denotes invariance under poor initialisations and accuracy of convergence denotes the compliance of the estimated mode with the global mode of the multi-modal similarity function. In the following a quantitative metric is derived based on the modelling of the mean shift tracking as random process, which allows for an analysis of the above mentioned criteria.

A. *Random Process Modelling*

The target object state χ_k to be estimated by the mean shift tracking algorithm is application-specific and may comprise the object location, orientation or scale. Basically, it can be modelled by the following linear system and measurement equations:

$$\chi_k = \chi_{k-1} + \mathbf{n}_k \tag{7}$$

$$\mathbf{y}_k = \chi_k + \mathbf{e}_k \tag{8}$$

The hidden object state χ_k follows from an unknown state transition which is modelled by an additive system noise process $\mathbf{n}_k$. Thus we are able to model the state uncertainty caused by the object's motion or deformation, for example. A measurement $\mathbf{y}_k$ of the state χ_k is obtained from the result of the mean shift tracking algorithm and yields a state estimation which ideally resembles the state χ_k. However, as the measurement depends on the mean shift initialisation, a measurement noise process $\mathbf{e}_k$ is introduced representing the error of the mean shift tracking caused by poor initialisations.

The basic idea is to derive a quantitative performance metric from the unknown distribution of the measurement noise process $\mathbf{e}_k$ estimated by a Monte Carlo simulation, which is driven by a predefined system noise process $\mathbf{n}_k$. Thereby a set of a priori particles $\{\hat{\chi}^-_{k,i}\}$ is predicted from an initial object state $\hat{\chi}^+_{k-1}$ according to the system equation (7):

$$\hat{\chi}^-_{k,i} = \hat{\chi}^+_{k-1} + \mathbf{n}_{k,i} \tag{9}$$

Note, that the set of particles $\{\mathbf{n}_{k,i}\}$ represents the predefined system noise process $\mathbf{n}_k$. Each a priori particle $\hat{\chi}^-_{k,i}$ initialises the mean shift tracking, yielding a measurement particle $\mathbf{y}_{k,i}$. Hence the mean shift tracking can be interpreted as non-linear mapping $f(\cdot)$:

$$\mathbf{y}_{k,i} = f(\hat{\chi}^-_{k,i}) \tag{10}$$

That way, we obtain a set of measurement particles $\{\mathbf{y}_{k,i}\}$ approximating the distribution of the measurement process $\mathbf{y}_k$. As we are interested in the measurement noise process $\mathbf{e}_k$, we determine the most likely object state $\hat{\chi}^+_k$ by the element-wise median of the set of measurement particles $\{\mathbf{y}_{k,i}\}$:

$$\hat{\chi}^+_k = \text{median}(\{\mathbf{y}_{k,i}\}) \tag{11}$$

Hence, the measurement noise process $\mathbf{e}_k$ can be approximated by the set of particles $\{\mathbf{e}_{k,i}\}$:

$$\mathbf{e}_{k,i} = \mathbf{y}_{k,i} - \hat{\chi}^+_k \tag{12}$$

The course of the above described Monte Carlo simulation is illustrated in Fig. 3.

ISBN: 978-1-61208-265-3

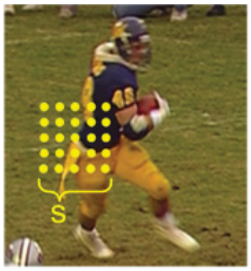


Figure 4. Multiple tracking initialisations.

The random process modelling for mean shift tracking described in this Section effectively resembles a multi-hypotheses tracking approach at which the mean shift tracking algorithm is evaluated for multiple hypotheses drawn from a distribution centred at an initial estimation. This approach is closely related to particle filtering [16] but combined with mean shift tracking.

B. Modelling of System Noise

The system noise process $\mathbf{n}_k$ steers the above described Monte Carlo simulation and consequently has an impact on the estimated measurement noise process $\mathbf{e}_k$. Particularly, the set of system noise particles $\{\mathbf{n}_{k,i}\}$ controls the quality of the mean shift initialisations $\{\hat{\chi}_{k,i}^{-}\}$ (compare (9)). Thus a large spread of the system noise process increases the occurrence of poor initialisations, leading to erroneous tracking results. For the sake of reproducibility, the system noise particles $\mathbf{n}_{k,i}$ are drawn from a deterministic, zero-mean process, which is derived by regular sampling of a cube with edge length (range) s. The resulting mean shift initialisations are exemplarily illustrated in Fig. 4. That way different noise processes $\{\mathbf{n}_k^s\}$ can be created by varying the parameter s, each resulting in a different measurement noise process $\mathbf{e}_k^s$.

C. Performance Metric

The distribution of the measurement noise process $\mathbf{e}_k^s$ is approximated by N_e particles $\mathbf{e}_{k,i}^s$. Hence these particles can be used to derive a metric for the evaluation of the tracking performance. We therefore use the mean absolute distance MAD_k^s :

$$\mathrm{MAD}_k^s = \mathcal{E}\{|\mathbf{e}_k^s|\} = \frac{1}{N_\mathrm{e}}\sum_{i=1}^{N_\mathrm{e}}|\mathbf{e}_{k,i}^s| \tag{13}$$

The defined metric can be used to evaluate the convergence behaviour of different object models for mean shift tracking, i.e. the convergence accuracy and convergence robustness. Possible experiments include the evaluation of the tracking performance for a fixed system noise process $\mathbf{n}_k^s$ across all frames $k \in \{1,\ldots,K\}$ of a test sequence or the evaluation for a set of system noise processes $\{\mathbf{n}_k^s\}_{s=1}^{S}$ and averaging the MAD_k^s over all frames k:

$$\mathrm{MAD}_s = \frac{1}{K}\sum_{k=1}^{K}\mathrm{MAD}_k^s \tag{14}$$

The latter approach allows an investigation of the robustness towards poor initialisations. Thus a larger value of the parameter s leads to an increased spread of the system noise process, which in turn increases the occurrence of poor initialisations.

IV. EVALUATION

We have implemented a mean shift algorithm for location tracking. The maximum number of mean shift iterations is set to 20 and the convergence bound is set to 0.1 pixels. As recommended in [5], the shadow kernel $K(\mathbf{x})$ is implemented by an Epanechnikov kernel whose bandwidth is equal to the dimension of the target object. Background colour information is not exploited and no update of the target model is performed during the course of tracking.

The presented evaluation results are obtained from three test sequences described by table I and Fig. 7 at which the accuracy and robustness of location tracking is evaluated with regard to the object representations presented in Fig. 2. All computations are based on $N_\mathrm{e} = 25$ initialisation samples, which are exemplarily illustrated in Fig. 4. For each sample, a complete sequence is processed. Thereby the initialisation in each frame, that is derived from the tracking result of the previous frame, is shifted according to the current sample. The resulting trajectories are then used to compute the value of MAD_s for each test sequence.

A. Test Sequences

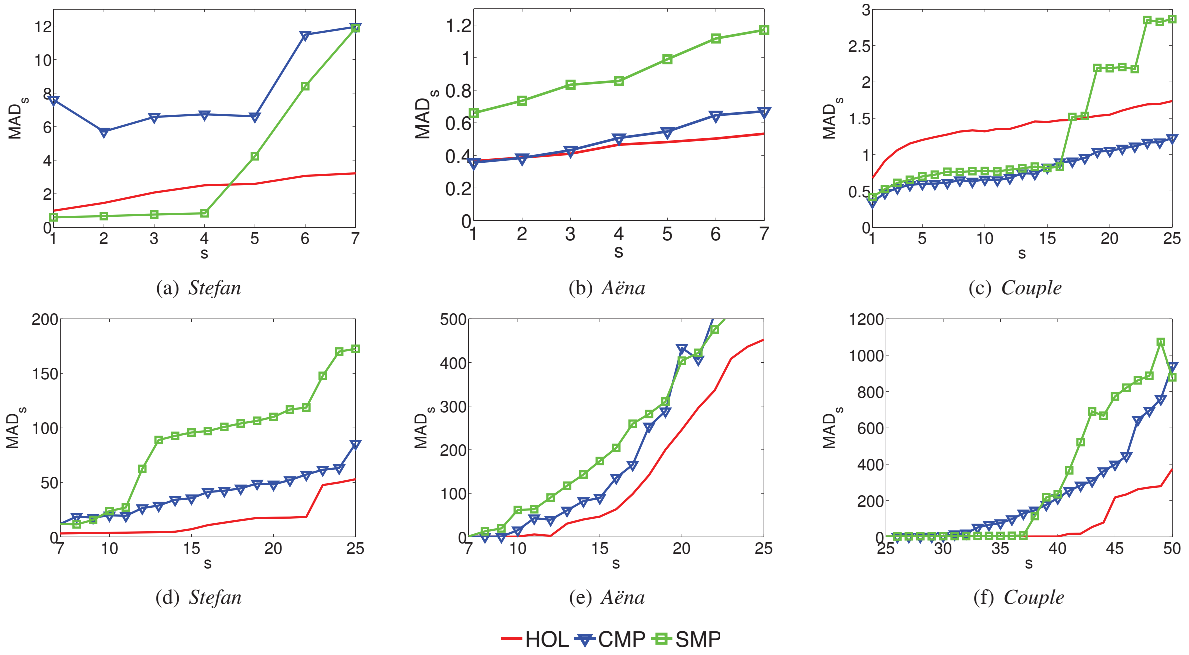
The test sequences feature different characteristics which affect the tracking performance. The *Stefan* sequence comprises tracking a tennis player against background clutter. A small and fast oscillating handbag of a lady is tracked in the *Aëna* sequence where the difficulty lies in the velocity of the target object. A much more distinct and easier target object is given by the pink dressed lady in the *Couple* sequence where, however, partial occlusion occurs.

B. Results

The values of MAD_s for all test sequences are plotted in Fig. 5 to assess the tracking performance across an entire test sequence for different ranges s of the initialisation region. In case of the *Stefan* sequence, the SMP object representation is superior to other object representations for small initialisation regions ($s < 5$) which is confirmed by

TABLE I. TEST SEQUENCES.

Sequence	Size	Target size	# frames
Stefan	355 × 288	70 × 177	300
Aëna	720 × 540	27 × 31	125
Couple	480 × 270	50 × 235	154



(a) *Stefan* (b) *Aëna* (c) *Couple*

(d) *Stefan* (e) *Aëna* (f) *Couple*

Figure 5. For different object representations the time-averaged MAD_k^s values have been evaluated for different ranges s of the initialisation region.

small MAD_s values indicating a low scatter of the mean shift tracking results (high accuracy).

This result is caused by a superior comprehension of spatial colour information by the SMP object representation in contrast to CMP or the holistic object representation. A drawback of both multi-part object representations is, however, given by a higher sensitivity (less robustness) to larger initialisation regions ($s > 5$).

An advantage of a multi-part object representation is, however, not apparent for the *Aëna* sequence. Fig. 5(b) and Fig. 5(e) demonstrate a superior accuracy and robustness of the holistic object representation for different ranges s of the initialisation region. This outcome is explained by the nearly uniform spatial colour distribution of the target object which allows no exploitation of spatial information by a multi-part object representation. Furthermore, the high velocity of the target object causes a high sensitivity to mean shift initialisations for all object representations which can be observed by the rapid increase of the MAD_s values in Fig. 5(e).

A more representative example for spatial colour information which can be exploited by multi-part object representations is given by the *Couple* sequence. For small initialisation regions ($s < 15$) both multi-part object representations yield a higher tracking accuracy proved by a small scatter of the mean shift tracking results as illustrated in Fig. 5(c) and 5(f). Due to the less distinctive background clutter, the multi-part object representations are more robust to poor mean shift initialisations in the *Couple* sequence than in case of the *Stefan* sequence. The robustness is more distinctive for the SMP object representation since it subdivides a target object only in vertical direction which is better suited for the target object of the *Couple* sequence.

For the sake of completeness, the temporal MAD_k^s is plotted for a fixed range of the initialisation region ($s = 10$) and a temporal segment of the *Couple* sequence in Fig. 6. This allows to identify key scenes for which certain mean shift object representations perform less accurate or less robust or which are more difficult for mean shift tracking in general. For example, the global peak in Fig. 6 corresponds to the period shortly after a partial occlusion where a more distinct scatter of the mean shift tracking results exists.

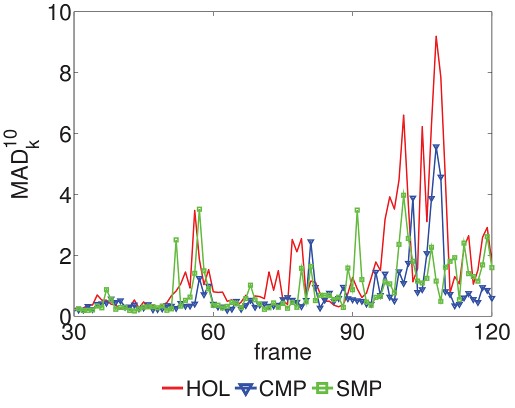


Figure 6. Progress of MAD_k^{10} values over time for the *Couple* sequence and different object representations.

ISBN: 978-1-61208-265-3

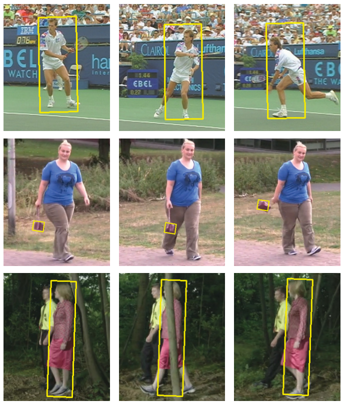


Figure 7. From top to bottom: Exemplary frames of the test sequences *Stefan*, *Aëna* and *Couple*.

V. CONCLUSION AND FUTURE WORK

We have introduced a novel approach for the quantitative evaluation of object representations in mean shift tracking, which does not rely on any ground truth data. Particularly, it has been demonstrated that the usage of multi-part object representations can improve the mean shift tracking accuracy. A drawback of multi-part object representations is, however, their higher sensitivity to poor mean shift initialisations which may occur during the tracking of highly agile target objects. A possible remedy is the combination of mean shift tracking with supportive algorithms, such as Kalman filter or Particle filter, which allow an initial prediction of the target object state.

The used object representation should, however, capture well the spatial colour distribution of the target object. Future work will therefore be focused on the development of an adaptive multi-part object representations that automatically adapts to the varying appearance of the object. As this online learning comprises the risk of a drift towards an invalid object representation a combination with a segmentation approach might also be promising.

ACKNOWLEDGEMENT

This work was co-funded by the German federal state North Rhine Westphalia (NRW) and the European Union (European Regional Development Fund: Investing In Your Future)

REFERENCES

[1] R. Mohan, J. Smith, and C. Li, "Adapting multimedia internet content for universal access," IEEE Transactions on Multimedia, vol. 1, no. 1, 1999, pp. 104–114.

[2] F. Pereira and I. Burnett, "Universal multimedia experiences for tomorrow," Signal Processing Magazine, IEEE, vol. 20, no. 2, 2003, pp. 63–73.

[3] S. De Bruyne, P. Hosten, C. Concolato, M. Asbach, J. De Cock, M. Unger, J. Le Feuvre, and R. Van de Walle, "Annotation based personalized adaptation and presentation of videos for mobile applications," Multimedia Tools and Applications, vol. 55, no. 2, 2011, pp. 307–331.

[4] A. Yilmaz, O. Javed, and M. Shah, "Object tracking: A survey," ACM Computing Surveys (CSUR), vol. 38, no. 4, 2006, pp. 1–45.

[5] D. Comaniciu, V. Ramesh, and P. Meer, "Kernel-based object tracking," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 25, no. 5, 2003, pp. 564–577.

[6] E. Maggio and A. Cavallaro, "Multi-part target representation for color tracking," in Proceedings of IEEE International Conference on Image Processing (ICIP'05), vol. 1, 2005, pp. 729–732.

[7] V. Parameswaran, V. Ramesh, and I. Zoghlami, "Tunable kernels for tracking," in Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR'06), vol. 2, 2006, pp. 2179–2186.

[8] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, J. Garofolo, R. Bowers, M. Boonstra, V. Korzhova, and J. Zhang, "Framework for performance evaluation of face, text, and vehicle detection and tracking in video: Data, metrics, and protocol," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 31, no. 2, 2009, pp. 319–336.

[9] C. J. Needham and R. D. Boyle, "Performance evaluation metrics and statistics for positional tracker evaluation," in Proceedings of International Conference on Computer Vision Systems (ICVS'03), 2003, pp. 278–289.

[10] D. Caulfield and K. Dawson-Howe, "Evaluation of multi-part models for mean-shift tracking," in Proceedings of International Machine Vision and Image Processing Conference (IMVIP'08), 2008, pp. 77–82.

[11] Y. Cheng, "Mean shift, mode seeking, and clustering," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 17, no. 8, 1995, pp. 790–799.

[12] D. Comaniciu and P. Meer, "Mean shift: A robust approach toward feature space analysis," IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 24, no. 5, 2002, pp. 603–619.

[13] G. Bradski, "Real time face and object tracking as a component of a perceptual user interface," in Proceedings of 4th IEEE Workshop on Applications of Computer Vision (WACV'98), 1998, pp. 214–219.

[14] L. Wang, C. Pan, and S. Xiang, "Mean-shift tracking algorithm with weight fusion strategy," in Proceedings of IEEE International Conference on Image Processing (ICIP'11), 2011, pp. 473–476.

[15] A. Yilmaz, "Kernel-based object tracking using asymmetric kernels with adaptive scale and orientation selection," Machine Vision and Applications, vol. 22, no. 2, 2011, pp. 255–268.

[16] S. Arulampalam, S. Maskell, N. Gordon, and T. Clapp, "A tutorial on particle filters for on-line non-linear/non-gaussian bayesian tracking," IEEE Transactions on Signal Processing, vol. 50, no. 2, 2002, pp. 174–188.