Deep motion features for visual tracking
Susanna Gladh, Martin Danelljan, Fahad Shahbaz Khan and Michael Felsberg

The self-archived postprint version of this conference article is available at Linköping University Institutional Repository (DiVA):
http://urn.kb.se/resolve?urn=urn:nbn:se:liu:diva-137896

N.B.: When citing this work, cite the original publication.

Original publication available at:
https://doi.org/10.1109/ICPR.2016.7899807
Copyright: Institute of Electrical and Electronics Engineers (IEEE)
http://www.ieee.org/
©2016 IEEE. Personal use of this material is permitted. However, permission to reprint/republish this material for advertising or promotional purposes or for creating new collective works for resale or redistribution to servers or lists, or to reuse any copyrighted component of this work in other works must be obtained from the IEEE.
Deep Motion Features for Visual Tracking

Susanna Gladh, Martin Danelljan, Fahad Shahbaz Khan, Michael Felsberg
Computer Vision Laboratory, Department of Electrical Engineering, Linköping University, Sweden

Abstract—Robust visual tracking is a challenging computer vision problem, with many real-world applications. Most existing approaches employ hand-crafted appearance features, such as HOG or Color Names. Recently, deep RGB features extracted from convolutional neural networks have been successfully applied for tracking. Despite their success, these features only capture appearance information. On the other hand, motion cues provide discriminative and complementary information that can improve tracking performance. Contrary to visual tracking, deep motion features have been successfully applied for action recognition and video classification tasks. Typically, the motion features are learned by training a CNN on optical flow images extracted from large amounts of labeled videos.

This paper presents an investigation of the impact of deep motion features in a tracking-by-detection framework. We further show that hand-crafted, deep RGB, and deep motion features contain complementary information. To the best of our knowledge, we are the first to propose fusing appearance information with deep motion features for visual tracking. Comprehensive experiments clearly suggest that our fusion approach with deep motion features outperforms standard methods relying on appearance information alone.

I. INTRODUCTION

Generic visual tracking is the problem of estimating the trajectory of a target in a sequence of images. It is challenging since only the initial position of the target is known. The various applications of generic tracking range from surveillance to robotics. Most state-of-the-art approaches follow the tracking-by-detection paradigm, where a classifier or regressor is discriminatively trained to differentiate the target from the background. Recently, Discriminative Correlation Filter (DCF) based methods [1], [2], [3], [4], [5] have shown excellent performance for visual tracking. These approaches efficiently train a correlation filter to estimate the classification confidences of the target. This is performed by considering all cyclic shifts of the training samples and exploiting the FFT at the training and detection steps. In this work, we base our approach on the DCF framework.

DCF based trackers typically employ hand-crafted appearance features, such as HOG [2], [4], Color Names [6], or combinations of these features [7]. Recently, deep Convolutional Neural Networks (CNNs) has been successfully applied for tracking [5], [8], [9], [10]. A CNN applies a sequence of convolution, normalization, and pooling operations on the input RGB patch. The parameters of the network are trained using large amounts of labeled images, such as the ImageNet dataset [11]. Deep convolutional features from pre-trained networks have been shown to be generic [12], and therefore also applicable for visual tracking.

Besides deep RGB features, deep motion features have been successfully employed for action recognition [13], [14]. These motion features are constructed by learning a CNN that takes optical flow images as input. The network is trained using flow data extracted from large amounts of labeled videos, such as the UCF101 dataset [15]. Unlike deep RGB networks, these deep flow networks capture high-level information about the motion in the scene. To the best of our knowledge, deep motion features are yet to be investigated for the problem of visual tracking.

Tracking methods solely based on appearance information struggle in scenarios with, for example, out-of-plane rotations (figure 1 first row), background distractors with similar appearance (figure 1 second row), and distant or small objects (figure 1 third row). In these cases, high-level motion cues provide rich complementary information that can disambiguate the target. While appearance features only encode static information from a single frame, deep motion features integrate information from the consecutive pair of frames used for estimating the optical flow. Motion features can therefore capture the dynamic nature of a scene that is complementary to appearance features. This motivates us to investigate the fusion of standard appearance features with deep motion features for visual tracking.

Contributions: In this paper, we investigate the impact of deep motion features for visual tracking. We use a deep optical flow network that was pre-trained for action recognition. Our approach does not require any additional labeled data for training the network. We investigate fusing hand-crafted and deep appearance features with deep motion features in a state-of-the-art DCF-based tracking-by-detection framework [3]. To show the impact of motion features, we further evaluate the fusion of different feature combinations.

Fig. 1. A comparison of using combined appearance information (hand-crafted HOG and deep RGB features) (in green) and our fusion of appearance and deep motion features (in red). Tracking results are shown for three example sequences: Box, Coupon and Skiing. Our fusion approach (red), using deep motion features, achieves superior results in these scenarios where appearance alone is insufficient.
Extensive experiments are performed on the OTB-2015 [16], Temple-Color [17], and VOT-2015 [18] datasets. On OTB-2015, our fusion of appearance and deep motion features significantly improves the baseline method, employing only appearance features, by 3.4% in mean overlap precision. Our fusion approach is further shown to advance the state-of-the-art performance with an absolute gain of 6.8% and 5.8% in mean overlap precision on OTB-2015 and Temple-Color respectively. Figure 1 shows a comparison of our fusion with deep motion features with the baseline method (employing only appearance features).

II. RELATED WORK

Discriminative tracking methods [3], [5], [19], [20] typically train a classifier or regressor for the task of differentiating the target from the background. These methods are often also termed tracking-by-detection approaches since they apply a discriminatively trained detector. Learning is performed online by extracting and labeling samples from the video frames. The training samples are often represented by e.g. raw image patches [1], [21], image histograms and Haar features [19], color [5], [20], or shape features [2], [4]. Lebeda et al. [22] use edge-points and estimate the transformation between each frame to successfully perform tracking on texture-less objects. Céhovin et al. [23] employ a couple-layer visual model, which combine local and global target appearance, to perform robust tracking of objects that have rapid appearance changes.

Among tracking-by-detection approaches, the Discriminative Correlation Filter (DCF) based trackers have recently shown excellent performance on standard tracking benchmarks [18], [24]. The key for their success is the ability to efficiently utilize limited data by including all shifts of local training samples in the learning. DCF-based methods train a least-squares regressor for predicting the target confidence scores by utilizing the properties of circular correlation and the fast Fourier transform (FFT). The MOSSE tracker [1] first considered training a single-channel correlation filter based on grayscale image samples of target and background appearance. A remarkable improvement is achieved by extending the MOSSE filter to multi-channel features. This can be performed by either optimizing the exact filter for offline learning applications [25] or using approximative update schemes for online tracking [2], [4], [6].

Despite their success, the DCF approaches are affected by the periodic assumption of the training samples, leading to negative boundary effects and a restricted training and search region size. This problem has recently been addressed in [3] by adding a spatial regularization term determined by the penalty weight $\alpha$. In the SRDCF formulation, the aim is to train a multi-channel convolution filter $f$ consisting of one $M \times N$ filter $f^l$ per feature dimension $l$. The target confidence scores for an $M \times N$ feature map $x$ are computed as $S_f(x) = \sum_{l=1}^{d} x^l \cdot f^l$. Here, $\ast$ denotes circular convolution.

To learn the filter $f$, the SRDCF formulation minimizes the squared error between the confidence scores $S_f(x_k)$ and the corresponding desired scores $y_k$.

$$\varepsilon(f) = \sum_{k=1}^{t} \alpha_k \| S_f(x_k) - y_k \|^2 + \sum_{l=1}^{d} \| w \cdot f^l \|^2. \quad (1)$$

The weights $\alpha_k$ determine the impact of each training sample and $\cdot$ denotes point-wise multiplication. The SRDCF employs a spatial regularization term determined by the penalty weight $w$. The filter is trained by minimizing the least squares loss (1) in the Fourier domain using iterative sparse solvers. We refer to [3] for more details about the training procedure. Target detection is explained further in section V.
IV. VISUAL FEATURES

The visual feature representation is a core component of a tracking framework. In this work, we investigate the use of a combination of hand-crafted features, deep appearance features and deep motion features for tracking.

A. Hand-crafted features

Hand-crafted features are typically used to capture low-level information, such as shape, color or texture. The Histograms of Oriented Gradients (HOG) is popularly employed for both visual tracking [2], [3], [4] and object detection [28]. It mainly captures shape information by calculating histograms of gradient directions in a spatial grid of cells. The histogram for each cell is then normalized with respect to neighboring cells to add invariance.

Other than shape features, various color-based feature representations have been commonly employed for tracking. For instance, the use of simple color transformations [20], [29] or color histograms [30]. Recently, the Color Names (CN) descriptor have been popularly employed for tracking due to its discriminative power and compactness [6]. The CN descriptor applies a pre-learned mapping from RGB to the probabilities of 11 linguistic color names.

B. Deep features

Features extracted by a trained Convolutional Neural Network (CNN) are known as deep features. The CNN consists of a number of computational layers that perform convolution, local normalization, and pooling operations on the input image patch. The final layers are usually fully connected (FC) and include an output classification layer. CNNs are typically trained in a supervised manner on large datasets of labeled images, such as ImageNet [11].

Feature representations learned by CNNs trained for a particular vision problem (e.g. image classification) have been shown to be generic and can be applied for a variety other vision tasks. For this purpose, most works apply the activations from the FC layer [31], [32]. Recently, activations from the convolutional layers have shown improved results for image classification [33], [34]. The deep convolutional features are discriminative and possess high-level visual information, while preserving spatial structure. Convolutional activations at a specific layer from a multi-channel feature map, which can be directly integrated into the SRDCF framework. Shallow-layer activations encode low-level features at a high spatial resolution, while deep layers contain high-level information at a coarse resolution.

Deep RGB Features: For the RGB images we use the imagenet-vgg-verydeep-16 network [31], with the MatConvNet library [35]. This network contains 13 convolutional layers. We investigate using both a shallow and a deep convolutional layer. For the shallow RGB layer, we use the activations from the fourth convolutional layer, after the Rectified Linear Unit (ReLU) operation. It consists of 128 feature channels and has a spatial stride of 2 pixels compared to the input image patch. For the deep layer of the RGB network, we use activations at the last convolutional layer, again after the ReLU-operation. This layer consists of a 512-dimensional feature map with a spatial stride of 16 pixels. Figure 2 shows example activations from the shallow layer (first row) and deep layer (second row) of the RGB network.

Deep Motion Features: The motion features are extracted using the approach described by [27]. We start by calculating the optical flow for each frame, together with the previous frame, according to [36]. The motion in the x- and y-directions forms a 3-channel image together with the flow magnitude. The values are adjusted to the interval [0, 255]. For our experiments, we use the pre-trained optical flow network provided by [13]. It is pre-trained on the UCF101 dataset [15] for action recognition and contains five convolutional layers. For the motion network, we only use the activations from the deepest convolutional layer. Similar to the RGB network, we extract the activations after the ReLU-operation. The resulting feature map consists of 384 channels at a spatial stride of 16 pixels compared to the input. An example optical flow image is displayed in figure 2, along with corresponding shallow (third row) and deep (fourth row) activations from the motion network.

V. OUR TRACKING FRAMEWORK

Here, we describe our tracking framework where we investigate the fusion of hand-crafted and deep appearance features with deep motion features. Our framework is based on learning an independent SRDCF model for each feature map. That is, we learn one filter $f_j$ for each feature type $j$. In a frame $k$, we extract new training samples $x_{j,k}$ for each feature type $j$ from the same image region centered at the estimated target location. We use a quadratic training region with an area equal to 5$^2$ times the area of the target box. For example, in our final tracking approach we combine three different feature maps: HOG $x_{1,k}$, the deep RGB layer $x_{2,k}$ and the deep motion layer $x_{3,k}$ (see section VI-A). The fused feature maps have different dimensionalities $d_j$ and spatial resolutions, leading to a different spatial sample size $M_j \times N_j$ for each feature $j$. The label function $y_{j,k}$ for feature $j$ is set to an $M_j \times N_j$
Convolution property, the DFT coefficients of the confidence maps have the DFT coefficients obtained from each filter to a pixel-dense grid. We then fuse scores obtained from each filter to be fused. The fused confidence scores are computed as,

\[ s = \frac{1}{J} \sum_{j=1}^{J} \mathcal{F}^{-1} \left\{ \mathcal{P}_{R \times S} \left\{ \hat{S}_j(z_j) \right\} \right\} \tag{2} \]

We obtain pixel-dense confidence scores (2) of the target by setting \( R \times S \) to be the size (in pixels) of the image region used for extracting the feature maps \( x_{j,k} \). The new target location is then estimated by maximizing the scores \( s(m,n) \) over the pixel locations \((m,n)\). To also estimate the target size, we apply the filters at five scales with a relative scale factor of 1.02, similar to [3], [7].

### VI. Experiments


#### A. Baseline Comparison

We investigate the impact of deep motion features by evaluating different combinations of appearance and motion representations on the OTB-2015 dataset. Table I shows a comparison of different feature combinations using mean overlap precision (OP) and area-under-the-curve (AUC). OP is defined as the percentage of frames in a video where the intersection-over-union overlap exceeds a certain threshold. In the tables, we report the OP at the threshold of 0.5, which corresponds to the PASCAL criterion. The AUC score is computed from the success plot, where OP is plotted over the range of thresholds [0, 1]. See [16] for further details about the evaluation metrics.

<table>
<thead>
<tr>
<th></th>
<th>HOG</th>
<th>RGB(s)</th>
<th>RGB(d)</th>
<th>RGB(s+d)</th>
<th>HOG+RGB(s)</th>
<th>HOG+RGB(d)</th>
<th>HOG+RGB(s+d)</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Mean OP (%)</strong></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Without deep motion features</td>
<td>74.5</td>
<td>74.1</td>
<td>56.3</td>
<td>78</td>
<td>74.9</td>
<td>80.7</td>
<td>79.1</td>
</tr>
<tr>
<td>With deep motion features</td>
<td>81.3</td>
<td>81</td>
<td>58.9</td>
<td>81.1</td>
<td>79.5</td>
<td><strong>84.1</strong></td>
<td><strong>82.2</strong></td>
</tr>
</tbody>
</table>

### Table I: Comparison of different combinations of hand-crafted (HOG), deep RGB and deep motion features on the OTB-2015 dataset. Results are reported in terms of mean overlap precision (OP) and area-under-the-curve (AUC) in percent. The two best results are displayed in red and blue font respectively. Shallow and deep layers of the RGB network are denoted RGB(s) and RGB(d). For each combination of appearance features, we also report the results when including deep motion features. We omit the result of shallow motion features, since we did not observe a gain in performance when including them. The fusion with deep motion features significantly improves the performance for all combinations.
We validate our approach in a comprehensive comparison with 12 state-of-the-art trackers: Struck [19], CFLB [26], LCT [37], ACT [6], KCF [5], DSST [2], SAMF [7], DAT [38], MEEM [20], HCF [5], SRDCF [3] and SRDCFdecon [39]. We also compare with the DeepSRDCF [9], that employs the shallow layer of a deep RGB network in the SRDCF.

Table II (first row) presents a state-of-the-art comparison, in mean OP, on the OTB-2015 dataset. The HCF tracker employing an ensemble of deep RGB features obtains a mean OP score of 65.5%. The SRDCF tracker using hand-crafted appearance features achieves a mean OP score of 72.9%. The DeepSRDCF employs appearance based RGB features and obtains a mean OP score of 77.3%. Our approach that combines hand-crafted and deep appearance based features with deep motion features achieves state-of-the-art results on this dataset with an overlap of 84.1%. Figure 3a presents the success plot for top-10 trackers on the OTB-2015 dataset. The area-under-the-curve (AUC) for each method is shown in the legend of the plot. The SRDCF obtains an AUC score of 54.3%. Our approach obtains state-of-the-art results by significantly outperforming the DeepSRDCF tracker with a gain of 3%.

C. Temple-Color Dataset

Next, we validate our proposed tracker on the challenging Temple-Color dataset [17]. The dataset consists of 128 videos. Table II (second row) presents a state-of-the-art comparison in mean OP. The HCF tracker obtains a mean OP score of 58.2%. The SRDCF tracker with hand-crafted appearance features provides a mean OP score of 62.2%. The DeepSRDCF further improves the results and obtains a mean OP score of 65.4%. Our approach obtains state-of-the-art results on this dataset with a mean OP of 71.2%. A significant gain of 5.8% in mean OP is obtained over the DeepSRDCF tracker. Figure 3b presents the success plot for top-10 trackers on the Temple-Color dataset. The area-under-the-curve (AUC) for each tracker is shown in the legend of the plot. The SRDCF obtains an AUC score of 51.6%. Among the existing methods, the DeepSRDCF provides the best results and achieves an AUC score of 54.3%. Our approach obtains state-of-the-art results by significantly outperforming the DeepSRDCF tracker with a gain of 3%.

D. VOT2015 Dataset

Table III presents a state-of-the-art comparison on the VOT2015 dataset [18] in comparison to the top 10 participants in the challenge according to the VOT2016 rules (see http://votchallenge.net). The dataset consists of 60 challenging videos compiled from a set of more than 300 videos. Here, the performance is measured in terms of accuracy (overlap with the ground-truth) and robustness (failure rate). The proposed method yields superior accuracy compared to the previously most accurate method (RAJSSC) and superior robustness compared to the previously most robust method (EBT). Compared to previous SRDCF-based methods, both accuracy and robustness are significantly improved.

VII. CONCLUSIONS

We investigate the impact of deep motion features in a DCF-based tracking framework. Existing approaches are limited to using either hand-crafted or deep appearance based features. We show that deep motion features provide complementary information to appearance cue and their combination leads to significantly improved tracking performance. Experiments are performed on three challenging benchmark tracking datasets: OTB-2015 with 100 videos, Temple-Color with 128 videos, and VOT2015 with 60 videos. Our results clearly demonstrate that fusion of hand-crafted appearance, deep appearance and deep motion features leads to state-of-the-art performance on both datasets. Future work includes incorporating deep motion features into the recent C-COT [8] framework to enable joint fusion of the multi-resolution feature map.
### Acknowledgments

This work has been supported by SSF (CUAS), VR (EMC2), FFI (CCAIS), WASP, NSC and Nvidia.

### References


<table>
<thead>
<tr>
<th>STTracker</th>
<th>RAJSSC</th>
<th>Struck</th>
<th>NSAMF</th>
<th>SC-EBT</th>
<th>pST</th>
<th>LDP</th>
<th>SRDCF</th>
<th>EBT</th>
<th>DeepSRDCF</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td>Robustness</td>
<td>1.77</td>
<td>1.63</td>
<td>1.26</td>
<td>1.29</td>
<td>1.86</td>
<td>1.48</td>
<td>1.84</td>
<td>1.24</td>
<td>1.02</td>
<td>1.05</td>
</tr>
<tr>
<td>Accuracy</td>
<td>0.52</td>
<td>0.37</td>
<td>0.47</td>
<td>0.53</td>
<td>0.55</td>
<td>0.51</td>
<td>0.56</td>
<td>0.47</td>
<td>0.47</td>
<td>0.58</td>
</tr>
</tbody>
</table>

### Table III

**Comparison with State-of-the-Art Methods on the VOT2015 Dataset. The Proposed Method Provides State-of-the-Art Accuracy and Robustness.**

### Fig. 4

Attribute-based analysis of our approach on the OTB-2015 dataset. Success plots are shown for 4 attributes. For clarity, we show the top 10 trackers in each plot. The title of each plot indicates the number of videos labeled with the respective attribute. Our approach provides consistent improvements compared to state-of-the-art methods on all 11 attributes.